tvst

Article

Automated Identification of Referable Retinal Pathology in Teleophthalmology Setting

Qitong Gao[1], Joshua Amason[2], Scott Cousins[2], Miroslav Pajic[1,3,*], and Majda Hadziahmetovic[2,*]

[1] Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
[2] Department of Ophthalmology, Duke University, Durham, NC, USA
[3] Department of Computer Science, Duke University, Durham, NC, USA

Correspondence: Majda Hadziahmetovic, Department of Ophthalmology, Duke University, 2351 Erwin Rd, Durham, NC 27710, USA. e-mail: majda.hadziahmetovic@duke.edu

Received: September 25, 2020
Accepted: January 31, 2021
Published: May 25, 2021

Keywords: deep learning; automated diagnosis; retinal pathology; image analysis; teleophthalmology

Citation: Gao Q, Amason J, Cousins S, Pajic M, Hadziahmetovic M. Automated identification of referable retinal pathology in teleophthalmology setting. Transl Vis Sci Technol. 2021;10(6):30, https://doi.org/10.1167/tvst.10.6.30

Purpose: This study aims to meet a growing need for a fully automated, learning-based interpretation tool for retinal images obtained remotely (e.g. teleophthalmology) through different imaging modalities that may include imperfect (uninterpretable) images.

Methods: A retrospective study of 1148 optical coherence tomography (OCT) and color fundus photography (CFP) retinal images obtained using Topcon's Maestro care unit on 647 patients with diabetes. To identify retinal pathology, a Convolutional Neural Network (CNN) with dual-modal inputs (i.e. CFP and OCT images) was developed. We developed a novel alternate gradient descent algorithm to train the CNN, which allows for the use of uninterpretable CFP/OCT images (i.e. ungradable images that do not contain sufficient image biomarkers for the reviewer to conclude absence or presence of retinal pathology). Specifically, a 9:1 ratio to split the training and testing dataset was used for training and validating the CNN. Paired CFP/OCT inputs (obtained from a single eye of a patient) were grouped as retinal pathology negative (RPN; 924 images) in the absence of retinal pathology in both imaging modalities, or if one of the imaging modalities was uninterpretable and the other without retinal pathology. If any imaging modality exhibited referable retinal pathology, the corresponding CFP/OCT inputs were deemed retinal pathology positive (RPP; 224 images) if any imaging modality exhibited referable retinal pathology.

Results: Our approach achieved 88.60% (95% confidence interval [CI] = 82.76% to 94.43%) accuracy in identifying pathology, along with the false negative rate (FNR) of 12.28% (95% CI = 6.26% to 18.31%), recall (sensitivity) of 87.72% (95% CI = 81.69% to 93.74%), specificity of 89.47% (95% CI = 83.84% to 95.11%), and area under the curve of receiver operating characteristic (AUC-ROC) was 92.74% (95% CI = 87.71% to 97.76%).

Conclusions: Our model can be successfully deployed in clinical practice to facilitate automated remote retinal pathology identification.

Translational Relevance: A fully automated tool for early diagnosis of retinal pathology might allow for earlier treatment and improved visual outcomes.

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has brought teleophthalmology into the spotlight and highlighted the need for a well-run remote retinal imagining model that, besides good image quality, provides an accurate and fast image interpretation.

This project is a part of the large initiative to perform retinal screening in patients with diabetes during their visits to the primary care provider's office. This paper focuses on our efforts to develop an automated system that can efficiently process

retinal images obtained during these visits and identify patients who need further ophthalmology attention.

Several groups have attempted to address this issue by proposing automated solutions that are either human-in-the-loop systems or operated semi-autonomously. However, developing a fully automated approach was challenging as a significant percentage of uninterpretable images were present in training and testing datasets.[1–23] Uninterpretable images exist due to inappropriate focus, exposure, or illumination settings used during the image-capturing process and do not contain sufficient image biomarkers for the reviewer to conclude the absence or presence of retinal pathology (i.e. ungradable).[24] Specifically, computer-aided diagnosis tools developed by Usher et al.[5] were able to identify retinal pathology in a semi-automated manner using color fundus photography (CFP) images. However, human interaction was necessary for the image preprocessing or feature extraction steps. Gargeya et al.[2,3,7,8,18,19] improved the automation degree, but for interpretable images only, by proposing a one-fit-for-all preprocessing method for CFP images, with the resulting images being processed and classified by convolutional neural networks (CNNs). Additionally, Kermany et al.[9,14,15,17,20] devised CNN-based models that can identify ophthalmic pathologies from optical coherence tomography (OCT) scans. To further improve the prediction performance and capture the image features jointly across different modalities, Yoo et al.[10–13] proposed multistream CNN models for automated diagnosis using multimodal inputs (e.g. OCT and CFP). However, to the best of our knowledge, no existing work can be deployed for fully automated retinal pathology diagnosis, mostly because uninterpretable images are excluded from training and testing.[2–21] This process requires an expert's input to determine ungradable images and exclude them from the dataset. The presence of images with substandard quality is universal and inevitable to encounter in clinical practice.[22,23] This problem might become more accentuated in the future with broader acceptance of automated image capture systems with integrated AI-based diagnosis algorithms. In such instances, no clinician would be present on-site to fine-tune the scanner for each patient or re-take images if the outputs were unsatisfactory. Consequently, a substantial number of ophthalmic screenings on undilated pupils will likely contain uninterpretable images, and it is essential to include those while designing such deep learning (DL) models to allow for their integration into a fully automated diagnosis system for instant and accurate diagnoses.

The purpose of this study was to create such an accurate DL approach for retinal image classification and identification of referable retinal pathology. Our main goal was to develop a CNN model that can automatically handle imperfect images, including uninterpretable images, and provide high validation accuracy and low false-negative rate to identify retinal pathology.

Materials and Methods

Retinal Imaging

This retrospective study analyzed 1148 OCT and CFP retinal images obtained from 647 patients with diabetes. Images were captured by Topcon, Maestro 3D-OCT multimodality OCT/Fundus imaging device (Topcon Inc., Tokyo, Japan). CFP had an angle of 45 degrees ± 5%, or 30 degrees, on the nondilated pupil. B scan horizontal range was 3–12 mm degrees ± 5%, with a 4× "Moving Average" oversampling performed, with the averaged final image. All eligible patients were invited to participate in the study and verbally consented to participate in the study by their primary care provider. The images were taken by trained certified medical assistants (CMAs). The study was a part of the Duke Quality Assessment/Quality Improvement (QA/QI) project and received institutional review broad approval from Duke University Health System. The study complied with the principles of the Declaration of Helsinki.

Dataset Formulation

Retinal images (OCT and CFP) were saved in JPEG compression format with a size of 659 × 512 and 661 × 653 pixels. For each OCT volume scan, only the central scan (i.e. the 31st B-scan out of a total of 60 B-scans in each volume scan) through the fovea was used. The images were resized to 299 × 299 to comply with the input dimension of the developed CNN architecture (for more details, see subsection: CNN Design). Images were graded as previously described (Hadziahmetovic et al., JAMA Ophthalmology)[24] by Duke medical retina fellows and a medical retina faculty, and the final grading of de-identified images was done by consensus. The images were classified as follows: (a) uninterpretable images (if no clear identification of macula was available due to poor positioning or inferior exposure owing to media opacity; containing 2 OCT images and 71 CFP images), (b) retinal pathology negative (RPN; containing 982 OCT and 952 CFP images), and (c) retinal pathology positive (RPP; containing 164 OCT and 125 CFP images; see Table 1). For each patient, there was at least one

Table 1. Distribution of the Original Dataset, Augmented Training Set, and Testing Dataset

	By Modality					
	OCT			CFP		
	Retinal Pathology Negative	Retinal Pathology Positive	Uninterpretable	Retinal Pathology Negative	Retinal Pathology Positive	Uninterpretable
Original	982	164	2	952	125	71
Augmented training set	3189	1736	14	2839	1302	798
Testing set	73	40	1	68	32	14

	By Eye		
	Retinal Pathology Negative	Retinal Pathology Positive	Total
Original	924	224	1148
Augmented training set	2601	2338	4939
Testing set	57	57	114

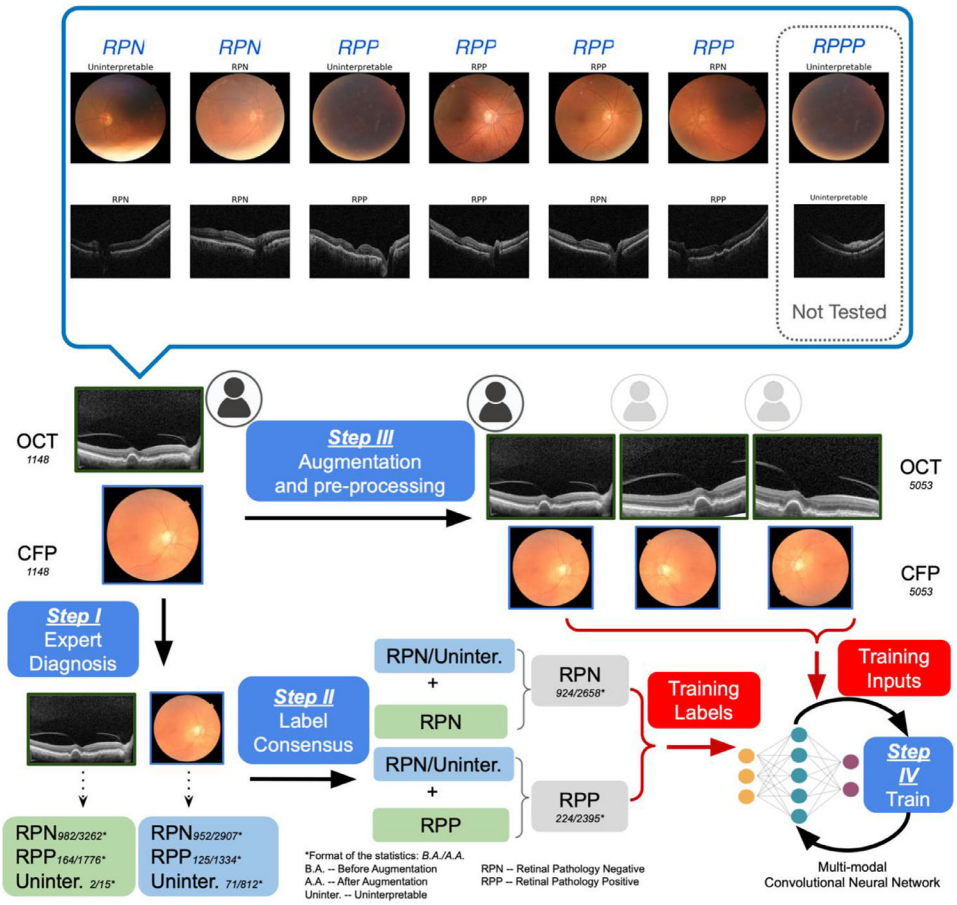

Figure 1. Overview of the proposed CNN model design methodology. The OCT and CFP images obtained from the automated screening system were first labeled respectively by experts (step I), and the individual diagnoses were used to generate training labels according to the Label Consensus Mechanism (step II). The two types of images were augmented and pre-processed to constitute the inputs to the CNN (step III), before being used, along with the obtained labels, for the CNN training (step IV).

interpretable image out of all obtained images. The final diagnosis used to train the CNN model was generated using the label consensus mechanism (LCM) presented in Appendix A and Supplementary Table S1. As a result, 924 eyes were labeled as normal (i.e. RPN), whereas 224 eyes were identified as RPP (Table 1 and Fig. 1). To form the testing dataset, we randomly selected 57 RPN and 57 RPP eyes from the available data following a uniform distribution. These numbers represented about 10% of the total eyes, roughly 6% and 25% of RPN and RPP eye cohorts, respectively. Uninterpretable images were present in 15 eyes (1 OCT

and 14 CFP). The remaining images were used to form the training dataset. We specifically chose this ratio of RPN and RPP samples to assure a well-balanced testing dataset and guarantee that the resulting dataset contained sufficient samples from the minority class (i.e. RPP).

Study Design and Outcomes Measures

We propose a fully automated system that utilizes a multimodal CNN to identify referable retinal pathology. Additionally, we propose a backpropagation algorithm associated with the CNN model that can train it to minimize the impact of the input images that do not contain sufficient biomarkers to determine diagnoses.

Problem Formulation

Pairs of OCT and CFP scans (O_k, and C_k) were obtained from each eye of each patient P_k, with some of them being uniterpretable. We designed a CNN model that takes input as (O_k, and C_k) and classifies it as "without" (i.e. RPN) and "with" (i.e. RPP) retinal pathology. Precisely, "without pathology" corresponds to the cases with normal OCT and CFP, and "with retinal pathology" refers to cases where retinal pathology can be identified in at least one of the imaging modalities (i.e. OCT or CFP). Moreover, if either O_k or C_k were uninterpretable, the outcome was derived from the interpretable image. Finally, if both O_k and C_k were uninterpretable, we specifically assigned the label as retina pathology potentially present (RPPP); those samples potentially could be selected and removed from the dataset using a separate classification model, as the clinicians would need to perform a further assessment, and potentially redo the imaging. (A detailed introduction of this labeling mechanism for paired OCT/CFP inputs is in Appendix A and Supplementary Table S1).

CNN Design

Design of CNN model was performed in three phases: (1) expert diagnosis and label consensus (steps I and II); (2) image augmentation and preprocessing (step III); and (3) training with the novel backpropagation algorithm that can work with uninterpretable images (step IV); illustrated in Figure 1.

Expert Diagnosis and Label Consensus (Steps I and II)

Each OCT and CFP image was individually labeled by the panel of retina professionals as uninterpretable, RPN, and RPP. Then, to train the CNN model, we determined the final diagnosis as RPN if one imaging modality was deemed uninterpretable and other RPN or both were RPN. Similarly, we labeled a patient RPP if we had at least one modality read as RPP. In the case of both modalities being uninterpretable, we referred to it as RPPP (Appendix A and Supplementary Table S1).

Image Augmentation and Preprocessing (Step III)

Bearing in mind that our dataset was limited (which is often the case with clinical data), we augmented the dataset by rotation, random cropping, flipping, etc.[25] Given that OCT images usually come with extensive background noise, which can prevent the DL-based models from capturing the image biomarkers,[26] we applied Gaussian filters[27] for noise reduction. No images were augmented for the validation set. However, the OCT images were de-noised using Gaussian blur, as in the training set. Details are introduced in Appendix B.

CNN Model Architecture and the Back Propagation Algorithm (Step IV)

We developed a multimodal CNN that takes as an input OCT and CFP images jointly and classifies them into RPN and RPP categories (Fig. 2). First, the input OCT and CFP images were processed by two sets of convolutional filters to obtain corresponding feature maps. Then, the output feature maps were fed into the global average pooling layers for dimension reduction (to derive feature vectors for both imaging modalities), which was then fed into a global, fully connected layer designed to: (1) map feature vectors to logits; and (2) to implicitly reach a consensus between the prediction outcomes (as the results from different imaging modalities could oppose each other – e.g. pathology does exist in one and does not in the other). Finally, Softmax activation was applied to the output layer to map the logits to probabilities of classifying the inputs as RPP. To ensure that the CNN can successfully handle uninterpretable images presented in both training and testing datasets, we developed an alternate gradient descent (AGD) algorithm. This way, we could minimize the impact of uninterpretable images on the prediction performance implicitly without formulating the binary classification problem as a multicategory task (e.g. RPN, RPP, and uninterpretable).

The AGD algorithm. We first divided all the weight parameters θ in the CNN into three subsets θ_1, θ_2 and θ_3, which represent the weights for the convolutional blocks and global average pooling layer that process the OCT inputs (i.e. Conv_blocks_1 and Avg_pool_1 in Fig. 2), the convolutional and global average pooling layers for the CFP modality (i.e. Conv_blocks_2 and Avg_pool_2 in Fig. 2), and the final fully connected layer (i.e. FC_3 in Fig. 2), respectively. The following briefly illustrates how the AGD algorithm works

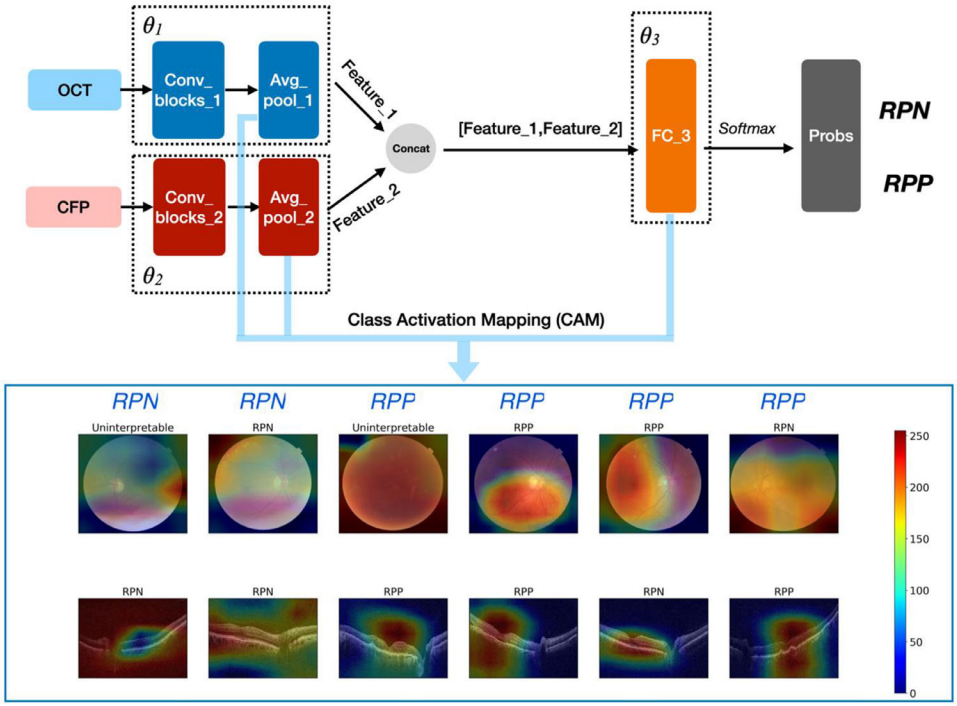

Figure 2. The architecture of the proposed CNN model with Class Activation Mapping (CAM). The OCT and CFP modalities are first processed with two sets of convolutional filters respectively; the resulting features are then concatenated and processed by a fully connected layer (θ_3) for classification. CAMs are generated using the outputs from the two global average pooling layers and weights from the fully connected layer.

during the training of the CNN model. In each training iteration, (I) we first updated θ_1 by minimizing the binary cross-entropy loss (BCEL) between the CNN predictions corresponding to the input (O_k, and C_k) samples that contain interpretable OCT images and the labels associated with them (i.e. in this step, the uninterpretable images were not included while calculating the training loss); (II) then similarly, θ_2 was updated by minimizing the BCEL between the CNN predictions corresponding to the input (O_k, and C_k) samples with interpretable CFP images from the training inputs and the labels associated with them; and (III) finally, θ_3 was updated to minimize the BCEL between the CNN predictions given all input (O_k, and C_k) samples (i.e. both interpretable and uninterpretable OCT/CFP) and the associated labels. After step I and II, the convolutional filters processing the OCT and CFP modality (i.e. θ_1, and θ_2) were trained toward extracting features that can best differentiate RPN/RPP samples if the inputs were interpretable. On the other hand, if one modality (or both modalities) of the inputs was (were) uninterpretable, then the features extracted by the corresponding convolutional filters were considered uninformative, as they were not included during the training of θ_1 and θ_2. In step III, the weights of the fully connected layer θ_3 were optimized to capture if the features output from θ_1 and θ_2 implies RPN, RPP, or uninformative, as well as learn to infer the correct predictions when the features corresponding to the OCT and CFP modality carry inconsistent information (e.g. one implies RPN whereas the other implies RPP, or the other was uninformative). As a result, the CNN was trained, using the AGD algorithm, to implicitly handle the uninterpretable images contained in the dual-inputs (O_k and C_k) without classifying them as a third class besides RPN and RPP. The illustration of the AGD algorithm from the mathematical perspective is provided in Appendix C (The Python code implementing this algorithm can be accessed from https://github.com/gaoqitong/Alternate-Gradient-Descent-For-Uninterpretable-Images).

Transfer learning. Transfer learning was applied to pretrain the convolutional blocks (i.e. θ_1, and θ_2) in the CNN model, as it was shown to be effective in boosting both training efficiency and validation performance.[28–30] Specifically, we used the open-source OCT dataset containing 108,312 OCT scans from 4 different categories: (1) choroidal neovascularization (37,206 images), (2) diabetic macular edema (DME; 11,349 images), (3) Drusen (8,617 images), and (4) normal (51,140 images), which were provided by

Kermany et al.[31] We also used the CFP image dataset containing 35,126 CFP images with (25,810 images) and without diabetic retinopathy (DR) pathology (9316 images), which are obtained from Kaggle.[32] Then, all the CFP images with DR pathology were flipped over horizontally and vertically to balance the number of images in the two classes, and loose pairing[33] was performed to couple the OCT and CFP modality, which then generated 100,000 "nominal" eyes. We further labeled the OCT images that contained any pathology as RPP and took a logical and between the individual OCT and CFP labels to determine the final diagnoses used to pretrain the network. Given that these two datasets did not contain any uninterpretable images, we pretrained the network, as illustrated in Figure 2, by minimizing the cross-entropy loss between the CNN predictions and labels for all images. Appendix C illustrated this optimization problem from a mathematical perspective. Although the open-source OCT dataset did not contain all retinal pathologies that we were interested in, the CNN model was still trained to effectively locate the biomarkers that help distinguish inputs as RPN and RPP, as presented in the Results section.

Specific convolutional layer architecture and training hyperparameters. The convolutional blocks in both the OCT and the CFP branches of the network (i.e. the Conv_blocks_1 and Conv_blocks_2 from Fig. 2) used the inception-v3[34] architecture. Furthermore, the open-source OCT dataset that we used to pretrain the CNN model had also been shown to attain the highest accuracy with the inception-v3 structure.[31] Specifically, during training, both OCT and CFP images were resized to 299 × 299 to comply with the design of the convolutional layers before feeding into the network.[34,35] After performing global average pooling for the OCT and CFP streams, the image features (i.e. Feature_1 and Feature_2 in Fig. 2) had the size nx1 × 1 × 2048, where n denotes the batch size. The two feature vectors were then concatenated and reshaped to an nx4096 vector, which was then processed by a fully connected layer with 4096 nodes to generate prediction logits. Finally, Softmax functions were applied to normalize the logits as probabilities of classifying the inputs as RPN/RPP. During training, Adam optimizer[28,29] was used to minimize training losses, where the learning rate was set to be 1e-04 with exponential decay of 0.91 in 1500 steps.

Results

To validate our approach, we selected the following three baseline methods to compare with our method: (1) training two CNN models that classifies the OCT and CFP modality respectively into three categories (RPN, RPP, and uninterpretable), then the final diagnoses are determined following the LCM illustrated in Appendix A and Supplementary Table S1; (2) first, two classifiers are trained to classify the interpretability for the OCT and CFP modality separately, followed by two CNN models that identify the presence of retinal pathology for interpretable OCT and CFP images respectively, with the final diagnoses being determined using the LCM; and (3) a two-stream CNN model based on the state-of-the-art multimodal ophthalmological image analysis methods developed by Wang et al.,[11–13] which uses the CNN architecture that does not consider any uninterpretable images, but is trained to minimize the cross-entropy loss with conventional backpropagation algorithms, instead of the AGD, as proposed in our work. In Appendix D, we illustrated the intuition of designing Baseline A and B and their implementation details.

Table 2 shows the performance comparison between our approach and the baseline methods in terms of accuracy, false-negative rate (FNR), recall (or true positive rate), specificity (or true negative rate), and area under the curve (AUC) of the receiver operating characteristic (ROC) curve; FNR is defined as:

$$\mathrm{FNR} = \frac{FN}{TP + FN} = 1 - Recall, \qquad (1)$$

with FN representing the false negatives and TP referring to the true positives. We chose FNR as one of the metrics because it evaluates the portion of the RPP patients who are falsely identified as RPN, or, in other words, the patients who have retinal pathology presented but failed to be recognized by the automated diagnosis system due to erroneous classifications. Our approach achieved 88.60% accuracy with 95% confidence interval (CI) of 82.76% to 94.43%, which outperforms all three baseline methods, as shown in Table 2. We also attained an FNR of 12.28% with 95% CI of 6.26% to 18.31% (or recall of 87.72% with 95% CI of 81.69% to 93.74%, which outperforms baseline A and B.

To address the fact that baseline C results in a lower FNR (and thus higher recall) than our model, we created the accuracy-FNR plots (Fig. 3A, the blue curve represents our approach, whereas the orange shows baseline C) showing how the accuracy and FNR change when different decision thresholds are applied to the probabilities output from the CNN (which can be interpreted as the confidence of classifying the input samples as RPP cases).[34] All the thresholds are sampled uniformly between 0.5 and 1, where the top-right end points of both curves correspond to the

Table 2. Performance Comparison Among Our Approach, Baseline A, Baseline B, and Baseline C on the Full Testing Dataset

	Accuracy/No. (%, 95% CI)	FNR/No. (%, 95% CI)	Recall/No. (%, 95% CI)	Specificity/No. (%, 95% CI)	AUC % (95% CI)	*P* Values
Our approach	**101 (88.60%, 82.76%–94.43%)**	**7 (12.28%, 6.26%–18.31%)**	**50 (87.72%, 81.69%–93.74%)**	**51 (89.47%, 83.84%–95.11%)**	**92.74% (87.71%–97.76%)**	**<0.001**
Baseline A	93 (81.58%, 74.46%–88.70%)	19 (33.33%, 24.68%–41.99%)	38 (66.67%, 58.01%–75.32%)	*55 (96.49%, 93.11%–99.87%)*	83.58% (76.11%–91.05%)	<0.001
Baseline B	81 (71.05%, 62.73%–79.38%)	23 (40.35%, 31.34%–49.36%)	34 (59.65%, 50.64%–68.66%)	47 (82.46%, 75.47%–89.44)	74.05% (64.96%–83.14%)	<0.001
Baseline C	91 (79.82%, 72.46%–87.19%)	*6 (10.53%, 4.89%–16.16%)*	*51 (89.47%, 83.84%–95.11%)*	40 (70.18%, 61.78%–78.57%)	87.32% (80.71%–93.93%)	<0.001

Performance comparison between our approach (alternate gradient descent with binary output), baseline A (2 single modal CNNs as 3-output task), baseline B (interpretability classifiers followed by 2 single modal CNNs as 2-output task), and baseline C (two-stream CNNs representing state-of-the-art methods for 2-modal image analysis) on the full testing dataset.†

†Statistics in italic correspond to better performance achieved by baselines than our approach, which are discussed in detail in the Results section. CIs for accuracy, FNR, Recall and Specificity were generated following the Wilson score interval.[47] CI for AUC computed following Hanley et al.[48]

P values generated by performing McNemar's test between the predictions and labels.

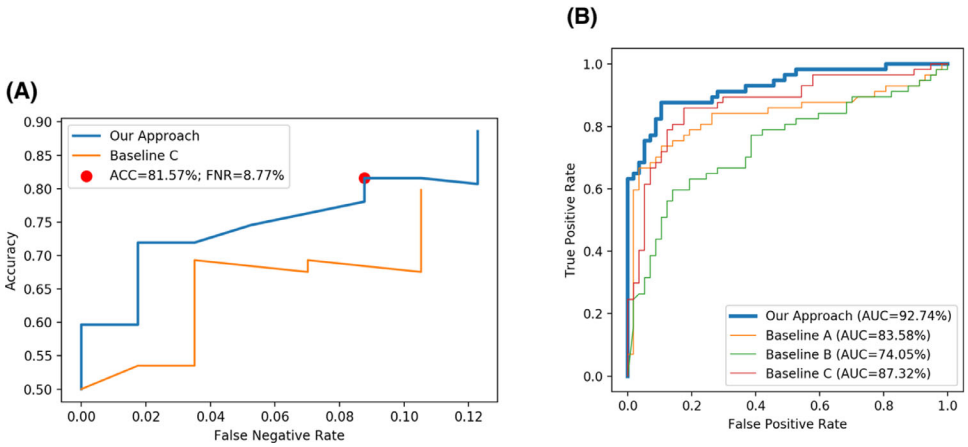

Figure 3. Accuracy-false negative rate (ACC-FNR) (**A**) curve and ROC (**B**) curve on the Testing Dataset. **A** ACC-FNR curve for our approach and baseline C. Baseline C has lower FNR than our approach with a decision threshold of 0.5; however, our method achieves both higher accuracy and lower FNR with a decision threshold providing optimal tradeoff between accuracy and FNR (e.g. the threshold of 0.65 as shown by the red dot in the plot). **B** ROC curves for our approach and baseline methods. Our approach achieves the highest AUC compared to all the baseline methods.

threshold of 0.5 (i.e. the samples that result in prediction probability greater than 0.5 are determined as RPP while the rest are classified as RPN) and the bottom-left points are associated with threshold 1 (i.e. all the inputs are classified as RPN regardless of the presence of pathology or not). As can be observed from Figure 3A, our method is capable of achieving an FNR of 8.77% with 95% CI of 3.58% to 13.96%, with an accuracy of 81.57% with 95% CI of 74.45% to 88.69%, which outperforms baseline C concerning both metrics given a threshold of 0.65 (shown as the red dot in Fig. 3A). Moreover, our method attains higher accuracy than baseline C for any decision threshold in [0.5 and , 1]. Our approach achieved a specificity of 89.47% with 95% CI of 83.84% to 95.11%, which outperforms baselines B and C. Note that baseline A gives rise to a higher specificity due to missclassifying RPP samples as RPN, which is indicated by the very high FNR (33.33%, 95% CI 24.68% to 41.99%) and the relatively low AUC (83.58%, 95% CI 76.11% to 91.05%). Finally, our approach reached an AUC of 92.74% with 95% CI of 87.71% to 97.76%, which is higher than all the baseline methods, as captured by the ROC curves shown in Figure 3B. Consequently, our approach achieved satisfactory performance evaluated through the five metrics and was able to balance between accuracy and FNR flexibly by selecting appropriate decision thresholds.

Table 3. Performance Comparison Among Our Approach, Baseline A, Baseline B, and Baseline C on the Dataset Containing Only Interpretable Images

	Accuracy/No. (%, 95% CI)	FNR/No. (%, 95% CI)	Recall/No. (%, 95% CI)	Specificity/No. (%, 95% CI)	AUC % (95% CI)	*P* Values
Our approach	**69 (88.46%, 81.37%–95.55%)**	**7 (16.67%, 8.40%–24.94%)**	**35 (83.33%, 75.06%–91.60%)**	**34 (94.44%, 89.36%–99.53%)**	**93.85% (88.39%–99.31%)**	**0.00766**
Baseline A	59 (75.64%, 66.11%–85.17%)	17 (40.48%, 29.58%–51.37%)	25 (59.52%, 48.63%–70.42%)	34 (94.44%, 89.36%–99.53%)	79.89% (71.00%–88.79%)	<0.001
Baseline B	65 (83.33%, 75.06%–91.60%)	10 (23.81%, 14.36%–33.26%)	32 (76.19%, 66.74%–85.64%)	33 (91.67%, 85.53%–97.80%)	89.48% (81.88%–97.09%)	<0.001
Baseline C	68 (87.18%, 79.76%–94.60%)	6 (14.29%, 6.52%–22.05%)	36 (85.71%, 77.95%–93.48%)	32 (88.89%, 81.91%–95.86%)	90.21% (83.31%–97.12%)	0.00443

Performance comparison between our approach (alternate gradient descent with binary output), baseline A (2 single modal CNNs as 3-output task), baseline B (interpretability classifiers followed by 2 single modal CNNs as 2-output task), and baseline C (two-stream CNNs representing state-of-the-art methods for 2-modal image analysis) on the dataset containing only interpretable images.

To evaluate the impact of the uninterpretable images on prediction performance, we evaluated our model by excluding them from the testing dataset (i.e. each eye with at least one uninterpretable image was excluded; Table 3). Performance of our model did not change when evaluated on interpretable images only. On the other hand, baseline B and C methods' performances increased dramatically in this setting, as expected, because both methods were not designed to process the uninterpretable inputs. Finally, baseline A had slightly decreased accuracy and recall, likely due to the higher FNR of the baseline A model. All of this could be observed by comparing the changes in the FNR between Tables 2 and 3, where the number of false negative samples barely decreased when the uninterpretable samples were excluded. In other words, for baselines B and C, by removing uninterpretable images, the classification performance improved, as those images lead to decreasing recall (or increasing FNR), while the opposite was true for baseline A. As presented, the uninterpretable images negatively impacted the baseline methods while having a minimal impact on our approach.

We further validated our model by generating class activation maps (CAMs), which can visualize how much "attention" the CNN model is paying to each pixel of the input images (see Fig. 2). We followed the procedure proposed by Zhou et al.,[36] where the weights of the fully connected layer (i.e. FC_3 in Fig. 2) and the image features generated from the global average pooling layers (i.e. Avg_pool_1 and Avg_pool_2 in Fig. 2) were used to generate attention values associated with all pixels in the input images from both imaging modalities. This evaluates to what extend each pixel is weighted while the CNN model generates predictions. The higher values correspond to the stronger attention, whereas they correspond lower to weaker attention (see Fig. 1).

Discussion

There is an unmet need for automated imaging and diagnosis systems for identifying retinal pathology. This limitation of the current healthcare model has been emphasized during the COVID-19 pandemic, especially because ophthalmology has been one of the hardest hit specialties.[37] Additionally, early recognition of sight-threatening retinal diseases might offer timely treatment, potentially improve visual outcomes, and reduce healthcare costs. Moreover, with improved triage, clinician effort, and clinic time might be better spent on other activities providing improved referral accuracy and more efficient use of ophthalmic resources.[38–41]

This paper introduces a CNN-based approach that enables fully automated retinal image classification into present or absent retinal pathology. Similar existing methods cannot be applied autonomously as they have not been developed while considering uninterpretable images, which are frequently encountered during eye screening,[2,3,7–19] and thus cannot handle them well. By addressing these limitations, our approach facilitates the development of automated retinal diagnosis systems, where a healthcare worker does not need to evaluate the quality of the images (in order for some to be retaken) before they are submitted for the analysis. This system can be deployed either in the clinics for triage or during remote screening (e.g. teleophthalmology) without involving physical interactions between patients and physicians.

Herein, we presented a CNN model that takes OCT and CFP images as dual-modal inputs and

predicts if the corresponding eye has retinal pathology (e.g. DR, DME, and age-related macular degeneration [AMD]). Our model was able to process imperfect/uninterpretable images resulting from the patient's poor positioning during the screening or inappropriate parameters[22,23] (e.g. focus, exposure, and illumination). Inputs obtained from uninterpretable images were utilized during the training through a novel backpropagation algorithm that can minimize the impact from images that do not contain sufficient image biomarkers to be determined as RPN/RPP during the training process. We created a fully automated retinal pathology diagnosis system (i.e. that requires no human interaction). To train and validate our model, we collected 1148 pairs of CFP and OCT images from 674 patients, where each pair pertains to a single eye of a patient. We used a 9:1 ratio to split the training and testing dataset. Finally, we attained a validation accuracy of 88.6%, recall/sensitivity of 87.7%, specificity of 89.5%, and AUC for ROC of 0.93. We presented the case, which only considers the dual-modal inputs (OCT and CFP); regardless, the proposed approach can be further extended to include other imaging modalities (e.g. fundus autofluorescence). Moreover, we observed that the performance of baseline methods could be negatively impacted when uninterpretable images are used for testing. On the other hand, the performance of our approach was not affected when evaluated with either full testing dataset or interpretable images only.

Significant work related to this topic was done by Yoo et al.,[10] Wang et al.,[11] Vaghefi et al.,[12] and Xu et al.[13] Specifically, in the Yoo method[8] pretrained VGG-19[42] was used to convert input OCT and CFP images into feature vectors, which were then classified as AMD and non-AMD by random forests. In this work, the pretrained CNN was applied for feature extraction without fine-tuning and potentially could have led to unsatisfactory performance.[43] Precisely, most of the pretrained models were trained with standard datasets (e.g. ImageNet[44]) that do not contain ophthalmic images and the resulting models were potentially not optimized for analysis of OCT or CFP inputs. The other mentioned methods proposed CNN models for the multimodal identification of retinal diseases. Wang et al.[11] and Xu et al.[13] developed two-stream CNNs to jointly analyze the OCT and CFP images. First, each modality was processed by the corresponding stream through convolutional filters and pooling layers for feature extraction using ResNet-18[45] or ResNet-50[45] architectures. Then, the two streams' output features were concatenated and fed into a fully connected layer for classification. Slightly different CNN architecture was applied in the Vaghefi[12] method. Each single-modal stream consisted of a few customized convolutional layers for initial processing, together with the outputs across streams that were combined through max-pooling followed by Inception-ResNet-V2[46] for further processing and classification. Despite being ground-breaking, these methods neither evaluate nor handle uninterpretable images, making them unsuitable for remote retinal image assessments where uninterpretable and low-quality images regularly occur.

To capture and generalize the ideas behind these four methods and emphasize the importance of uninterpretable image utilization, we combined them in the baseline C learning approach and compared it to our model. We concluded that although baseline C achieved slightly lower FNR, it attained 9.3% less accuracy than our method when the presented decision thresholds were used in our model (see Table 2). However, when the decision thresholds in our model were adjusted, our approach achieved both lower FNR and higher accuracy than baseline C, but with slightly lower accuracy than with our initial decision threshold – this highlights that by controlling decision thresholds, we were able to make a tradeoff between accuracy and FNR.

Furthermore, this underlined the importance of taking into account uninterpretable images during the training phase and showed that our AGD algorithm and the obtained CNN model could effectively handle uninterpretable images. In addition, we designed baseline A and B models to evaluate the prediction performance when the AGD backpropagation algorithm was not used, and the input images were classified into three categories (i.e. RPN, RPP, and uninterpretable), as opposed to the two-class problem addressed by our model trained by the AGD algorithm. Comparing these two methods to ours showed that our method had higher accuracy and lower FNR. The improved performance of our method and baseline C compared to baseline A and B methods confirm the strength of multimodal analysis, where the CNN models can effectively capture the correlation among different imaging modalities and make accurate predictions.

Finally, FNR is an important factor to consider while validating different image interpretation models because it is crucial not to miss pathology that can have serious consequences. As shown in the ACC-FNR Curve in Figure 3A, our CNN based approach allowed the users to balance the tradeoff between accuracy and FNR by customizing the decision thresholds (i.e. a threshold around

0.5 can be applied for attaining higher accuracy, whereas a threshold greater than 0.5 leads to lower FNR).

Conclusion

We have developed a fully automated retinal image interpretation system that outperformed other existing computational models. Our multimodal input approach used two inputs (CFP and OCT) to process and identify the presence of retinal pathology, but it is not limited to only these imaging modalities. The novel backpropagation algorithm that we proposed was able to utilize low-quality or uninterpretable images in the decision making process (about 6% of all photographs), and proved that it was minimally affected by it.

Limitations

This approach has limitations, and we will be addressing them as part of our future research. First, we can potentially improve the prediction performance with a dataset containing more balanced labels. Given the FNR of 12.28%, the CNN model may still classify RPP images as RPN. This is highlighted by the fact that regardless of the augmentation, the effective sample size in the RPP group is outnumbered by the effective sample size in the RPN group. Second, although the dataset contains a fairly sufficient number of uninterpretable CFP images, a limited number of uninterpretable OCT images was available. This leads to unequal distribution and may potentially influence the final outcome if the dataset contains a higher number of uninterpretable CFP images. Third, our dataset did not contain the samples where both imaging modalities were uninterpretable; thus, we could not demonstrate the model's performance in that setting. However, the implicit binary classification mechanism (i.e. the AGD algorithm) did not hinder this analysis if such data were available in the dataset. Specifically, the CNN model could still be trained to classify the inputs into two categories (i.e. (i) RPN and (ii) RPP or RPPP). The latter (i.e. RPP and RPPP) samples could be grouped into one class because both should be referred further. Moreover, the CNN model (see Fig. 2) could identify RPPP samples as the only difference between processing two uninterpretable modalities and one (or zero) uninterpretable modality was the fully connected layer θ_3 needed to learn to map uninformative features generated by both convolutional streams (i.e. θ_1 and θ_2) to its corresponding label while the other feature was uninformative. Given that the AGD algorithm would not interfere with this process during training, we expect that the fully connected layer can learn from such samples and inference properly; thus, our approach could process the inputs constituted by two uninterpretable modalities. On the other hand, if one prefers to refer the RPPP cases separately from the RPP cases (i.e. classify them into two separate classes), an additional classification model can be introduced to identify the RPPP samples from the dataset before our approach is applied. Finally, our model does not identify specific retinal pathology (e.g. DR, AMD, and DME) but instead classifies the images as retina pathology present or absent. The main focus of our future work will be resolving this challenge.

Acknowledgments

The authors thank Topcon Inc., and Greg Hoffmeyer for providing us with the unit, training, and imaging software Harmony.

Author Contributions

Concept and design: Gao, Pajic, and Hadziahmetovic.

Acquisition, analysis, or interpretation of data: Gao, Amason, Hadziahmetovic, and Pajic.

Drafting of the manuscript: Gao, Amason, Hadziahmetovic, and Pajic.

Critical revision of the manuscript for important intellectual content: Hadziahmetovic, Pajic, and Cousins.

Statistical analysis: Gao.

Obtained funding: Pajic and Hadziahmetovic.

Administrative, technical, or material support: Hadziahmetovic and Pajic.

Supervision: Hadziahmetovic and Pajic.

Disclosure: **Q. Gao**, None; **J. Amason**, None; **S. Cousins**, NotalVision (I); and Stealth (C), PanOptica (C), Merck Pharmaceuticals (C), and Clearside Biomedical (C); **M. Pajic**, None; **M. Hadziahmetovic**, None

* MH and MP had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1–29.
2. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124(7):962–969.
3. Raman R, Srinivasan S, Virmani S, Sivaprasad S, Rao C, Rajalakshmi R. Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. *Eye*. 2019;33(1):97–109.
4. Winder RJ, Morrow PJ, McRitchie IN, Bailie JR, Hart PM. Algorithms for digital image processing in diabetic retinopathy. *Comput Med Imaging Graph*. 2009;33(8):608–622.
5. Usher D, Dumskyj M, Himaga M, Williamson TH, Nussey S, Boyce J. Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening. *Diabetic Med*. 2004;21(1):84–90.
6. Mookiah M RK, Acharya UR, Chua CK, et al. Computer-aided diagnosis of diabetic retinopathy: a review. *Comput Biol Med*. 2013;43(12):2136–2155.
7. 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.
8. Singh RK, DMENet Gorantla R.: Diabetic macular edema diagnosis using hierarchical ensemble of CNNs. *PLoS One*. 2020:15(2):e0220677.
9. Wang Y, Zhang Y, Yao Z, Zhao R, Zhou F. Machine learning based detection of age-related macular degeneration (AMD) and diabetic macular edema (DME) from optical coherence tomography (OCT) images. *Biomed Opt Express*. 2016;7(12):4928–4940.
10. Yoo TK, Choi JY, Seo JG, Ramasubramanian B, Selvaperumal S, Kim DW. The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. *Med Biolog Eng Comput*. 2019;57(3):677–687.
11. Wang W, Xu Z, Yu W, et al. Two-stream CNN with loose pair training for multi-modal AMD categorization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 156–164). Springer, Cham, 2019.
12. Vaghefi E, Hill S, Kersten HM, Squirrell D. Multimodal retinal image analysis via deep learning for the diagnosis of intermediate dry age-related macular degeneration: a feasibility study. *J Ophthalmol*. 2020;2020:7493419.
13. Xu Z, Wang W, Yang J, et al. Automated diagnoses of age-related macular degeneration and polypoidal choroidal vasculopathy using bi-modal deep convolutional neural networks. *Br J Ophthalmol*, https://doi.org/10.1136/bjophthalmol-2020-315817.
14. Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y. Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images. *Transl Vis Sci Technol*. 2018;7(6):41–41.
15. Treder M, Lauermann JL, Eter N. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. *Graefes Arch Clin Exp Ophthalmol*. 2018;256(2):259–265.
16. Treder M, Lauermann JL, Eter N. Deep learning-based detection and classification of geographic atrophy using a deep convolutional neural network classifier. *Graefes Arch Clin Exp Ophthalmol*. 2018;256(11):2053–2060.
17. Kermany DS, Goldbaum M, Cai W, Valentim C, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122–1131.
18. 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):1–13.
19. Shibata N, Tanito M, Mitsuhashi K, Fujino Y, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. 2018;8(1):1–9.
20. Muhammad H, Fuchs TJ, De Cuir N, et al. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. *J Glaucoma*. 2017;26(12):1086.
21. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology*. 2019;126(4):513–521.
22. Bennett TJ. Maximizing quality in ophthalmic digital imaging. *J Ophthalmic Photogr.* 2009;31(1):32–39.

23. Liu S, Paranjape AS, Elmaanaoui B, Dewelle J, et al. Quality assessment for spectral domain optical coherence tomography (OCT) images. Proc SPIE Int Soc Opt Eng. 2009;7171:71710X.
24. Hadziahmetovic M, Nicholas P, Jindal S, Mettu PS, Cousins SW. Evaluation of a remote diagnosis imaging model vs dilated eye examination in referable macular degeneration. *JAMA Ophthalmol.* 2019;137(7):802–808.
25. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60.
26. Moreno-Barea FJ, Strazzera F, Jerez JM, Urda D, Franco L. Forward noise adjustment scheme for data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI) Conference* (pp. 728–734). IEEE; 2018. Bangaluru, India.
27. Gonzalez R, Woods R. *Digital Image Processing*. Boston, Massachusetts: Addison-Wesley Publishing Company; 2007.
28. Saad D. Online algorithms and stochastic approximations. *Online Learning*. 1998;5:6–3.
29. Kingma D, Ba JL. ADAM: A Method for Stochastic Optimization. *Proceedings of the 2015 International Conference on Learning Representations*. 2015.
30. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
31. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122–1131.
32. Diabetic Retinopathy Detection – Identify signs of diabetic retinopathy in eye images, https://www.kaggle.com/c/diabetic-retinopathy-detection/data. Updated February 17, 2015. Accessed July 12, 2020.
33. Wang W, Xu Z, Yu W, et al. October. Two-stream CNN with loose pair training for multi-modal AMD categorization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 156–164). Cham: Springer; 2019.
34. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826), Las Vegas, NV, 2016.
35. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
36. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929), 2016.
37. Mehrotra A, Chernew M, Linetsky D, Hatch H, Cutler D. The impact of the COVID-19 pandemic on outpatient visits: a rebound emerges. *To the* Point (blog), Commonwealth Fund. Available at: https://www.commonwealthfund.org//publications/2020apr/impact-/covid-19-outpatient-visits.
38. Sreelatha OK, Ramesh SV. Teleophthalmology: improving patient outcomes? *Clin Ophthalmol (Auckland, NZ)*. 2016;10:285.
39. Rathi S, Tsui E, Mehta N, Zahid S, Schuman JS. The current state of teleophthalmology in the United States. *Ophthalmology*. 2017;124(12):1729–1734.
40. Li B, Powell AM, Hooper PL, Sheidow TG. Prospective evaluation of teleophthalmology in screening and recurrence monitoring of neovascular age-related macular degeneration: a randomized clinical trial. *JAMA Ophthalmol*. 2015;133(3):276–282.
41. Michalak S, Hadziahmetovic M. Developments in teleophthalmology for diabetic retinopathy: diabetic eye disease is a prime target for remote diagnosis and management. *Retinal Physician*, 2020;17:39–43.
42. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. *arXiv preprint arXiv:1409.1556*.
43. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems* (pp. 3347–3357), 2019.
44. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE; 2009.
45. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778), 2016.
46. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. 2016. *arXiv preprint arXiv:1602.07261*.
47. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*. 1927;22(158):209–212.
48. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.