tvst

Artificial Intelligence

RobOCTNet: Robotics and Deep Learning for Referable Posterior Segment Pathology Detection in an Emergency Department Population

Ailin Song[1,2], Jay B. Lusk[1], Kyung-Min Roh[2], S. Tammy Hsu[2], Nita G. Valikodath[2], Eleonora M. Lad[2], Kelly W. Muir[2], Matthew M. Engelhard[3], Alexander T. Limkakeng[4], Joseph A. Izatt[5], Ryan P. McNabb[2], and Anthony N. Kuo[2,5]

[1] Duke University School of Medicine, Durham, NC, USA
[2] Department of Ophthalmology, Duke University, Durham, NC, USA
[3] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA
[4] Department of Emergency Medicine, Duke University, Durham, NC, USA
[5] Department of Biomedical Engineering, Duke University, Durham, NC, USA

Correspondence: Anthony N. Kuo, Department of Ophthalmology, Duke University, DUMC Box 3802, Durham, NC 27710, USA. e-mail: anthony.kuo@duke.edu

Received: September 26, 2023
Accepted: January 31, 2024
Published: March 15, 2024

Keywords: artificial intelligence; deep learning; robotics; optical coherence tomography; acute eye care

Citation: Song A, Lusk JB, Roh KM, Hsu ST, Valikodath NG, Lad EM, Muir KW, Engelhard MM, Limkakeng AT, Izatt JA, McNabb RP, Kuo AN. RobOCTNet: Robotics and deep learning for referable posterior segment pathology detection in an emergency department population. Transl Vis Sci Technol. 2024;13(3):12, https://doi.org/10.1167/tvst.13.3.12

Purpose: To evaluate the diagnostic performance of a robotically aligned optical coherence tomography (RAOCT) system coupled with a deep learning model in detecting referable posterior segment pathology in OCT images of emergency department patients.

Methods: A deep learning model, RobOCTNet, was trained and internally tested to classify OCT images as referable versus non-referable for ophthalmology consultation. For external testing, emergency department patients with signs or symptoms warranting evaluation of the posterior segment were imaged with RAOCT. RobOCTNet was used to classify the images. Model performance was evaluated against a reference standard based on clinical diagnosis and retina specialist OCT review.

Results: We included 90,250 OCT images for training and 1489 images for internal testing. RobOCTNet achieved an area under the curve (AUC) of 1.00 (95% confidence interval [CI], 0.99–1.00) for detection of referable posterior segment pathology in the internal test set. For external testing, RAOCT was used to image 72 eyes of 38 emergency department patients. In this set, RobOCTNet had an AUC of 0.91 (95% CI, 0.82–0.97), a sensitivity of 95% (95% CI, 87%–100%), and a specificity of 76% (95% CI, 62%–91%). The model's performance was comparable to two human experts' performance.

Conclusions: A robotically aligned OCT coupled with a deep learning model demonstrated high diagnostic performance in detecting referable posterior segment pathology in a cohort of emergency department patients.

Translational Relevance: Robotically aligned OCT coupled with a deep learning model may have the potential to improve emergency department patient triage for ophthalmology referral.

Introduction

A growing body of literature has applied machine learning and deep learning techniques to diagnose posterior segment diseases. Deep learning has been used to detect diabetic retinopathy, macular degeneration, and papilledema in both fundus photographs and optical coherence tomography (OCT) images, as well as to make referral recommendations in ophthalmology settings.[1–12] These approaches have the potential to increase efficiency in patient care, improve diagnostic accuracy, and expand access to critical screening services.[13,14]

One major limitation of utilizing deep learning for automated screening and evaluation is that deep learning algorithms generally require high-quality input data, which often necessitates the involvement of specialized personnel, such as ophthalmic photographers, to gather data.[15] Workforce limitations make implementing these technologies in many settings infeasible: for example, as of 2022, there are only 60,000 ophthalmic technicians in the United States, whereas the U.S. population was estimated to require between 72.9 and 142.6 million eye visits annually in 2015.[16,17] As the U.S. population ages and the prevalence of eye conditions increases with age, the growing population-level need for eye care will likely exacerbate the existing burden on the limited eye care workforce.[18,19]

As one potential solution to human resource limitations, our group recently developed a robotic device capable of capturing retinal optical coherence tomography images via autonomous alignment.[20] This robotically aligned optical coherence tomography (RAOCT) system has the potential to improve care in nonophthalmic settings where access to ophthalmic expertise and equipment is limited, such as the emergency department. Urgent eye conditions are common, and improper or delayed diagnosis can lead to permanent visual impairment.[21,22] Previous studies have shown that the accuracy of referral diagnosis from emergency physicians is low, especially for conditions affecting the posterior segment of the eye.[23,24] We recently demonstrated that the use of RAOCT images improved emergency physicians' sensitivity for retinal and optic nerve abnormalities from 0% to 69%.[25] The diagnostic performance may be further improved if this imaging modality is combined with a deep learning model capable of autonomous triage of the images, which would create a diagnostic image acquisition and interpretation pipeline novel to the literature.

In this study, we coupled RAOCT imaging with a deep learning model to classify images as referable to ophthalmology versus non-referable to ophthalmology. Furthermore, prospective study designs and external testing on data different than the data used for model development are essential to ensure optimal generalizability and reliability of deep learning models.[26,27] Therefore we prospectively evaluated this system in a diverse population of patients from the emergency department. Our objective was to determine whether such a system could differentiate referable versus non-referable posterior segment pathology among emergency department patients.

Methods

Study Design

This study was a training, internal testing, and external testing study designed to evaluate a deep learning model for classifying retinal OCT images as referable versus non-referable for ophthalmology consultation. Any pathology necessitating evaluation by an ophthalmologist was considered referable. Figure 1 shows our workflow diagram. In short, two publicly available OCT datasets and a set of images of ophthalmology clinic patients previously obtained with our RAOCT system were used for training and internal testing of the deep learning model. For external testing, we applied the model to a set of RAOCT images prospectively obtained from a cohort of emergency department patients. Institutional Review Board approval was obtained from Duke University. Written informed consent was obtained from all enrolled study participants after explanation of the nature and possible consequences of the study.

Reporting Guidelines

This manuscript is reported in accordance with the 2015 updated Standards for Reporting Diagnostic Accuracy statement. A checklist is provided in Supplementary Table S1.

Training and Internal Testing Datasets

Two publicly available OCT datasets (Kermany et al.[5] and Srinivasan et al.[28]) and a dataset previously acquired with our RAOCT system were used for deep learning model training and internal testing. Supplementary Figure S1 shows representative images from the three datasets. The dataset from Kermany et al.[5] contained 84,484 temporal-nasal foveal OCT B-scans (Spectralis OCT; Heidelberg Engineering, Heidelberg, Germany) obtained as part of routine clinical care at five institutions in the United States and China.[5] This dataset included images of normal retinas, as well as examples of choroidal neovascularization, macular edema, and drusen.

The dataset adapted from Srinivasan et al.[28] contained 1824 B-scans from volumetric OCT imaging (Spectralis OCT) obtained at three U.S. academic institutions for research, which included 15 volumes with dry age-related macular degeneration and 15 volumes with diabetic macular edema. Because pathologies were potentially located only in parts of the volumes, two ophthalmologists experienced in the interpretation of

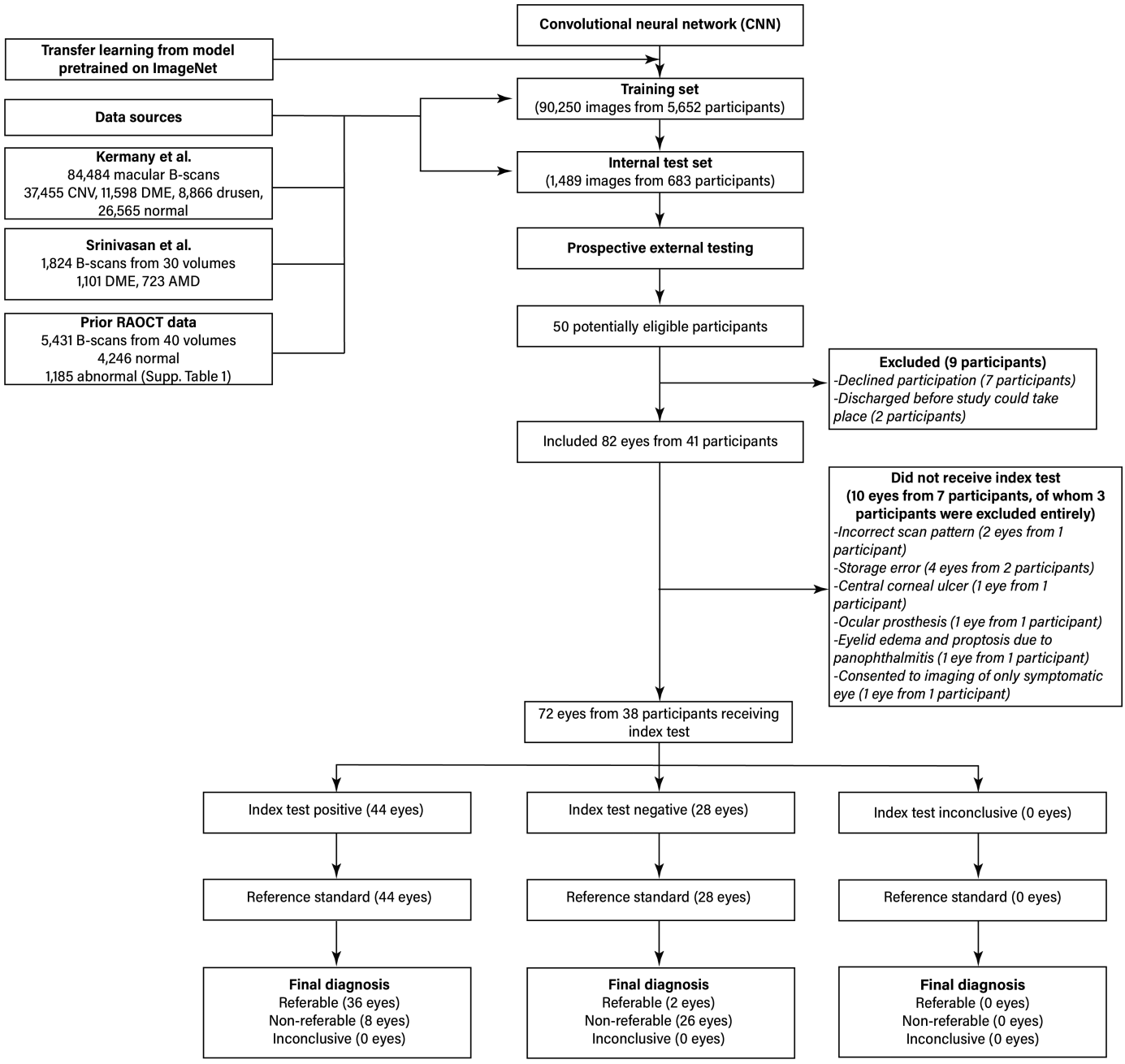

Figure 1. Study workflow diagram. A convolutional neural network pretrained on ImageNet data was retrained on two publicly available datasets[5,28] and a set of images previously obtained with our RAOCT system from ophthalmology clinic patients to create a deep learning model that classified OCT images as referable versus non-referable for ophthalmology consultation. A portion of these data were reserved for internal testing. We then externally tested the model on RAOCT images acquired from a cohort of emergency department patients with symptoms or signs that warranted evaluation of the posterior segment of the eye. For model evaluation, the primary outcome was AUC for detecting referable posterior segment pathology. AMD, age-related macular degeneration; CNV, choroidal neovascularization; DME, diabetic macular edema.

OCT images reviewed all B-scans of the volumes and excluded any B-scans without clear evidence of pathology from the dataset.

The dataset previously obtained with our RAOCT system included 5431 B-scans from volumetric OCT imaging of patients with a broad range of posterior eye diseases seen at the Duke Eye Center and subjects with normal retinas. Supplementary Table S2 describes the demographic and clinical characteristics of this dataset.

All posterior eye pathologies present in the three datasets above were considered referable to ophthalmology. In terms of data partition, the dataset from Kermany et al.[5] presented images readily divided into a training subset and a test subset. We used the images from their training subset for training and images from their test subset for internal testing. Both the dataset from Srinivasan et al.[28] and our own dataset were split randomly into training and internal testing subsets in a 14:1 ratio.

We additionally performed a sensitivity analysis excluding images derived from the RAOCT system from the training and internal test sets to assess whether including images obtained with the RAOCT system in training improved model performance in external testing on RAOCT images acquired in a different clinical setting.

Deep Learning Model

Our deep learning model used OCT B-scan images as its input and classified them as referable versus non-referable for ophthalmology consultation. Using Tensorflow,[29] we adapted an Inception V3 convolutional neural network pretrained on ImageNet data.[30] We used Inception V3 because it has previously been shown to have better performance than comparative convolutional neural networks such as ResNet-50 and VGG-19 and similar performance to ResNet-101,[31] and because Kermany et al.,[5] whose data make up the bulk of our training data, successfully used Inception V3 in their study. The layer "mixed6" was used to extract image features. The image features were then processed by a fully connected layer, a dropout layer, and a softmax output layer, which were the only trainable layers when retraining on the training set. Each input image was resized to 300 × 300 pixels to correspond to the default input size of the Inception V3 network and to maximize the tradeoff between batch size and image resolution.[32] The model parameters were trained with a RMSProp optimizer with a learning rate of 0.0001 in batches of 15 images. Training was set to run for 100 epochs with an additional early stopping procedure based on a threshold for training accuracy to avoid overfitting to the training set. Data augmentation was performed on the training set to increase the size and diversity of the dataset and to minimize bias arising from heterogeneity in imaging techniques. Data augmentation techniques included rotation, horizontal flip, width shift, height shift, shearing, and zooming. Manual hyperparameter search and optimization were performed. Training was conducted with an AMD Ryzen Threadripper 2950 × 16-Core central processing unit, using an NVIDIA RTX 3090 graphics processing unit, with 32 GB available in random access memory.

External Testing in an Emergency Department Population

For external testing, a cohort of adult patients presenting to the Duke University Hospital Emergency Department (ED), a Level I trauma center, between November 2020 and October 2021 were prospectively imaged with our RAOCT system after providing written consent. These patients presented with acute visual changes, headache, or focal neurologic deficit(s), and received an ophthalmology consult. Patients with hemodynamic instability, penetrating trauma to the eye, or inability to follow commands required for the OCT imaging procedure were excluded. Aside from these inclusion and exclusion criteria, patients were not excluded from enrollment or final analysis based on severity of pathology or ultimate diagnosis, or whether such a diagnosis can be made based on OCT imaging alone. The patients' medical records were reviewed for patient age at presentation, legal sex, and self-identified race. We performed an a priori sample size calculation on the patient level to develop a target for number of patients to enroll. To detect a minimum area under the curve (AUC) of 0.80 based on conventional thresholds and prior literature for the clinical application,[33–35] 34 patients would be needed to have 90% power at a significance level of 0.05.

The RAOCT platform used for imaging has been described previously.[20] In brief, we used a custom swept-source OCT system (100 kHz A-scan rate and $\lambda_0 = 1040$ nm) with 32° posterior eye field of view scanning optics mounted on a robot arm (Universal Robots UR3e). OCT volumes of 900 pixels × 250 pixels × 750 pixels (A-scans × B-scans × depth) were obtained. Both eyes of the patients were imaged unless an ocular condition limiting the utility of any ocular imaging was present. Patients sat freely in front of the system without chin or head rests, and the RAOCT system automatically located the eye of interest and maintained its alignment to the pupil (see Supplementary Video S1). A research technician was present for patient safety and to press a button to trigger the acquisition of the OCT scan. The technician had no clinical training and no experience or training working as an ophthalmic technician. The technician had training in human subject research, basic instruction in operation of the system including deployment of an emergency stop for the robotic system, and ready access to additional support if needed for patient safety. Clinical staff involved in patient enrollment were not

involved in image acquisition. Volumetric OCT images including both macular and optic nerve regions were acquired in less than two seconds. A single central B-scan for each eye was selected for external testing of the deep learning model. The model was only provided the central foveal B-scan of each patient. Deep learning image analyses occurred asynchronously after image capture and did not affect real-time patient care.

To establish the reference standard (diagnostic ground truth labels), we first retrospectively extracted from medical records the clinical diagnosis made by the ED consulting ophthalmologist based on examination. In addition, a retina specialist, who was masked to all patient data, reviewed the collected volumetric RAOCT images for presence or absence of referable pathology on imaging and noted any specific pathology present. If there was a disagreement between the ED consulting ophthalmologist's clinical diagnosis and the retina specialist's OCT assessment, a second senior retina specialist with extensive experience in OCT reading made the determination of whether a referable pathology was present or absent based on OCT and noted any specific pathology present.

In addition to evaluating model performance against the reference standard, two human experts (retina specialists) reviewed the central foveal B-scan images alone in the external testing set (i.e., without volumetric OCT or other clinical information) to classify each image as referable versus non-referable for ophthalmology consultation. This experiment was designed to compare the model versus human expert performance under equivalent testing conditions, as the model was only provided the central foveal B-scan images.

Statistical Analysis

The deep learning model was evaluated against ground truth labels for training, internal testing, and external testing. Receiver operating characteristic (ROC) and precision-recall (PR) curves were generated using classification probabilities of referable vs. non-referable. The primary performance metric was the AUC. The PR curves were summarized by average precision (AP). For external testing in emergency department patients, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were also calculated. These metrics were determined based on a point on the ROC curve selected to prioritize increased sensitivity, without unduly compromising specificity (i.e., the point where increasing sensitivity further would result in dramatic decreases in specificity), because the clinical application was to detect referable pathology. Generally, under-referral has more negative consequences for patients (i.e., missing a vision-threatening condition) than over-referral.[36] In a secondary analysis, to evaluate the model performance in comparison with human expert performance under comparable conditions (when human experts were provided with only the central foveal B-scan in the absence of any clinical information or volumetric OCT information), we calculated the human experts' sensitivity, specificity, PPV, and NPV for detecting referable posterior segment pathology. Bootstrapping was used to estimate 95% confidence intervals (CI) of the performance metrics. If the estimate of a given performance metric was 0% or 100%, the exact binomial method was used instead. Additionally, post-hoc subgroup analyses were performed for the most common conditions. Because the model and the human experts were only tasked with classifying images as referable versus non-referable and not tasked to identify specific conditions, only the true-positive count and the false-negative count for each condition can be deduced. Therefore, sensitivity was the only metric calculated in the analysis for specific conditions. Analysis was performed using Python (version: 3.8.5), and the "scikit-learn" library was used to calculate the performance metrics.[37]

Heatmap Generation

In an attempt to interpret the model's approach to image classification, the integrated gradients method was used to generate heatmaps showing areas contributing most to the model classification.[38] We used the integrated gradients method because, compared with class activation methods, the integrated gradients method has been shown to provide more accurate localization and coverage of pixels of importance to the model and because this method has previously been used for an interpretable artificial intelligence framework in ophthalmology.[14,38] The contributions were measured relative to a baseline black image intended to provide no information into the model. To visualize the underlying morphology, we overlaid the heatmaps on the OCT images as semitransparent color heatmaps. In general, heatmaps are more useful for images predicted to have referable pathology than images predicted to be non-referable, as in images predicted to be non-referable, the heatmap is the same for the prediction from the baseline (entirely black) image, as demonstrated in prior work.[14]

Results

We included 90,250 OCT images for training (59,967 referable and 30,283 non-referable) and 1489

images for internal testing (961 referable and 528 non-referable). As shown in Figure 1, for the external testing cohort, 50 patients were approached for enrollment. Seven patients declined participation, and two were discharged from the emergency department before study procedures could be completed; three further patients were excluded due to technical errors. Therefore a total of 72 eyes of 38 emergency department patients were included in the external testing set. Patient demographic and clinical characteristics as well as reference standard diagnoses for the external testing set are shown in Table 1. The median age was 58 (interquartile range, 37–66), and 47% were female. Thirty-eight (53%) had referable pathology, including a broad range of posterior segment pathologies such as retinal artery occlusion, optic nerve edema, retinal detachment, and acute retinal necrosis (Table 1). No adverse events occurred because of the performance of the index test (RobOCTNet) or the reference standard.

Figure 2 shows the ROC and PR curves for the training, internal testing, and external testing sets. In the internal test set, the model had an AUC of 1.00 (95% CI, 0.99–1.00) and an AP of 1.00 (95% CI, 1.00–1.00) for detection of referable posterior segment pathology. For external testing, the model had an AUC of 0.91 (95% CI, 0.82–0.97), an AP of 0.88 (95% CI, 0.76–0.98), a sensitivity of 95% (95% CI, 87%–100%), a specificity of 76% (95% CI, 62%–91%), a PPV of 82% (95% CI, 70%–92%), and an NPV of 93% (95% CI, 83%–100%).

We conducted post-hoc subgroup analyses for the most common conditions in the external testing set. Among eyes with optic nerve edema (n = 9) and eyes with epiretinal membrane (n = 10), our model had a sensitivity of 100% for referable posterior segment pathology. Among eyes with staphyloma or peripapillary atrophy (n = 10), the model had a sensitivity of 90%. Among the eyes with drusen (n = 11), the model had a sensitivity of 91%. Of note, in the single eye with drusen where the model failed to detect pathology, the drusen in question was not contained in the central foveal B-scan provided to the model.

Table 2 compares the performance of RobOCTNet with the performance of two human experts who reviewed the same central foveal B-scan images as RobOCTNet. In summary, RobOCTNet performed comparably to both experts. Of note, expert 2 considered the images for three (4%) eyes ungradable because of quality; for four (6%) additional eyes, expert 2 needed clinical history to decide. However, expert 1 was able to grade all images. Figure 3 compares the sensitivities of the model for the most common conditions with those of the human experts, demonstrating comparable or potentially superior performance of RobOCTNet.

Table 1. Demographic and Clinical Characteristics of Patients Included in the External Testing Set

Total Number of Patients	38
Total number of eyes	72
Median age, years (interquartile range)	58 (37-66)
Sex	
Male	20 (53%)
Female	18 (47%)
Race	
Caucasian/White	27 (71%)
Black/African American	7 (18%)
Asian	2 (5%)
Caucasian/White and Black/African American	1 (3%)
Not reported/declined	1 (3%)
Presenting symptoms/signs*	
Acute visual changes	36 (95%)
Headache	18 (47%)
Focal neurologic deficit(s)	2 (5%)
Non-referable per reference standard	34 (47%)
Referable per reference standard†	38 (53%)
Drusen	11 (15%)
Staphyloma or peripapillary atrophy	10 (14%)
Epiretinal membrane	10 (14%)
Optic nerve edema	9 (13%)
Optic atrophy	4 (6%)
Retinal artery occlusion	3 (4%)
Retinal vein occlusion	2 (3%)
Non-proliferative diabetic retinopathy	2 (3%)
Grade 2 hypertensive retinopathy	2 (3%)
Geographic atrophy	2 (3%)
Retinal detachment	1 (1%)
Acute retinal necrosis	1 (1%)
Chorioretinal lesion of unknown significance in the setting of contralateral endogenous panophthalmitis	1 (1%)
Intraretinal fluid in the setting of giant cell arteritis	1 (1%)

*Some patients had more than one sign/symptom, resulting in a sum greater than 100%.

†Some eyes had more than one diagnosis.

Heatmaps provided pixel-based maps that demonstrated the contribution of each pixel in the OCT images to predict the presence of referable pathology. Representative examples of these heatmaps are shown in Figure 4. In general, heatmaps highlighted areas of clinical importance in OCT images.

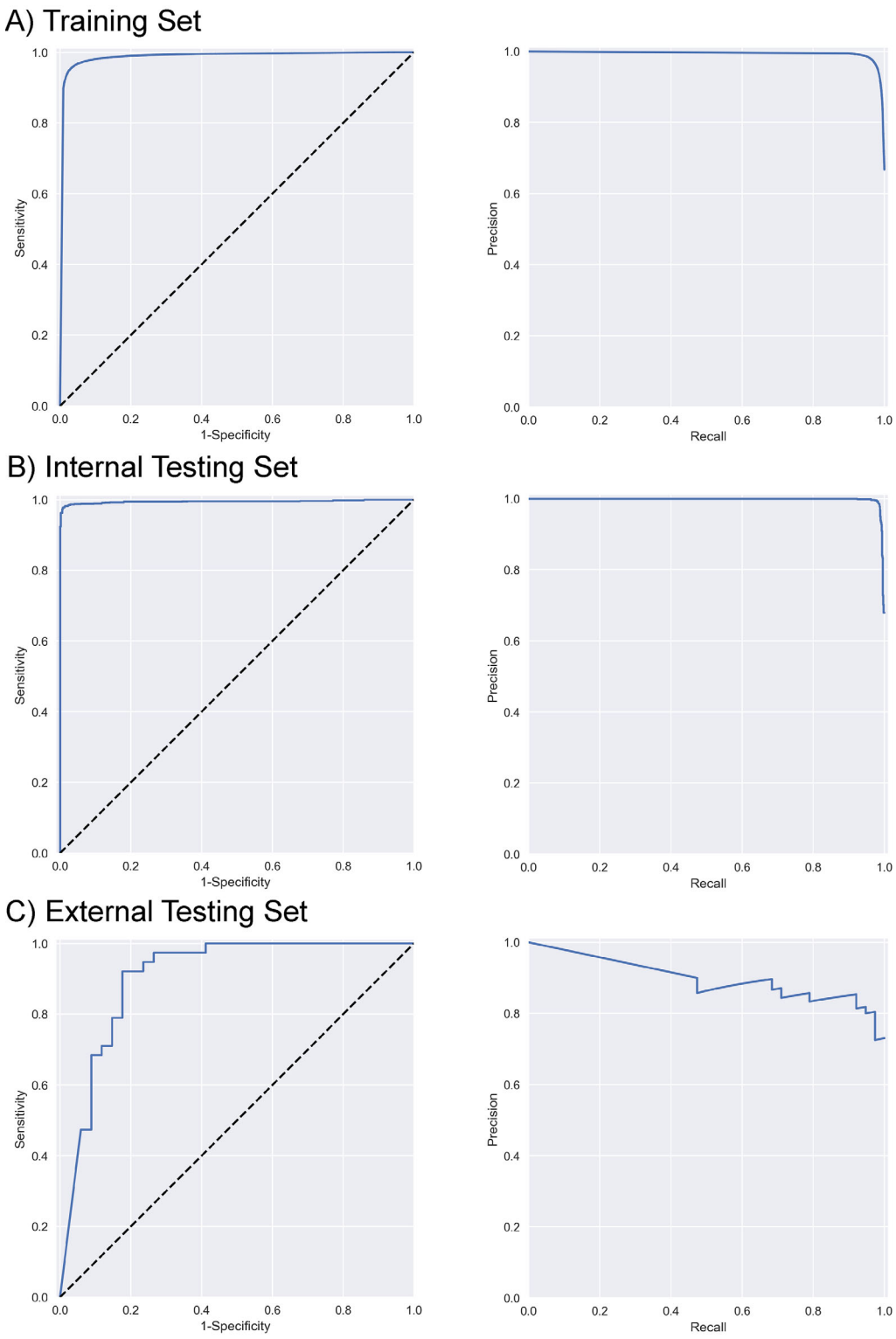

Figure 2. Receiver operating characteristic curves (*left*) and precision-recall curves (*right*) for the (**A**) training set, (**B**) internal testing set, and (**C**) external testing set. For detecting referable posterior segment pathology, the model had an AUC of 0.99 (95% CI, 0.99–0.99) and an AP of 0.99 (95% CI, 0.99–0.99) in the training set; an AUC of 1.00 (95% CI, 0.99–1.00) and an AP of 1.00 (95% CI, 1.00–1.00) in the internal testing set; and an AUC of 0.91 (95% CI, 0.82–0.97) and an AP of 0.88 (95% CI, 0.76–0.98) in the external testing set of emergency department patients.

Table 2. Performance Metrics of RobOCTNet Versus Human Expert Graders (Retina Specialists)

	Number of Eyes	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
RobOCTNet	72	95% (87%–100%)	76% (62%–91%)	82% (70%–92%)	93% (83%–100%)
Expert 1	72	68% (53%–82%)	100% (90%–100%)	100% (87%–100%)	74% (60%–86%)
Expert 2	65*	87% (74–97%)	44% (26–63%)	69% (54–81%)	71% (50–93%)
RobOCTNet	65*	95% (87–100%)	74% (57–89%)	84% (72–94%)	91% (78–100%)

*Expert 2 considered the images for three eyes ungradable because of quality and needed clinical history to make decisions about four additional eyes, so only 65 eyes were included in the analysis of expert 2's performance. We present the performance of RobOCTNet on both the full set of 72 eyes and the reduced set of 65 eyes for comparison with expert 2's performance.

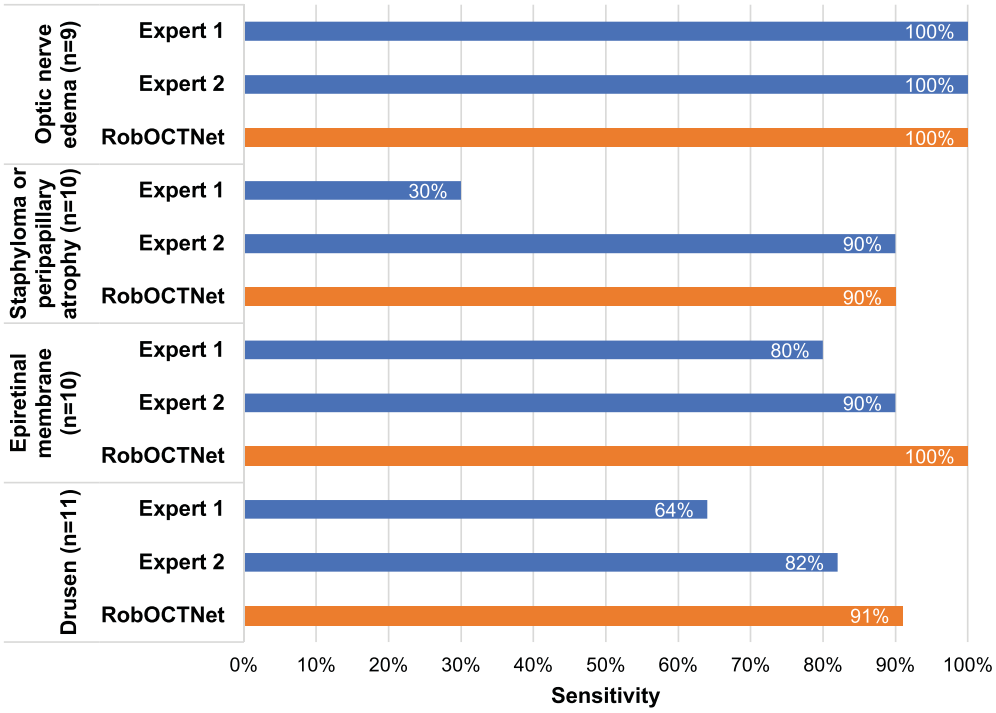

Figure 3. Bar graphs comparing the sensitivities of RobOCTNet with the sensitivities of two human experts (retina specialists) for the most common conditions in the external testing set.

In a sensitivity analysis where RAOCT images were excluded from the training and internal test sets, the model performed comparably in internal testing (AUC = 1.00; 95% CI, 1.00–1.00; AP = 1.00; 95% CI, 1.00–1.00). However, in external testing, the model did not perform as well as the model trained on RAOCT images: AUC was 0.85 (95% CI, 0.76–0.94), AP was 0.83 (95% CI, 0.68–0.95), sensitivity was 89% (95% CI, 80%–98%), specificity was 68% (95% CI, 51%–83%), PPV was 76% (95% CI, 63%–88%), and NPV was 85% (95% CI, 71%–97%).

Discussion

Our study coupled deep learning and robotics to identify referable optic nerve and retinal pathology for patients presenting to the emergency department with symptoms or signs warranting evaluation of the posterior segment. We showed that our deep learning/RAOCT system was effective at classifying presence versus absence of referable posterior segment pathology, with a sensitivity of 95% and a specificity of

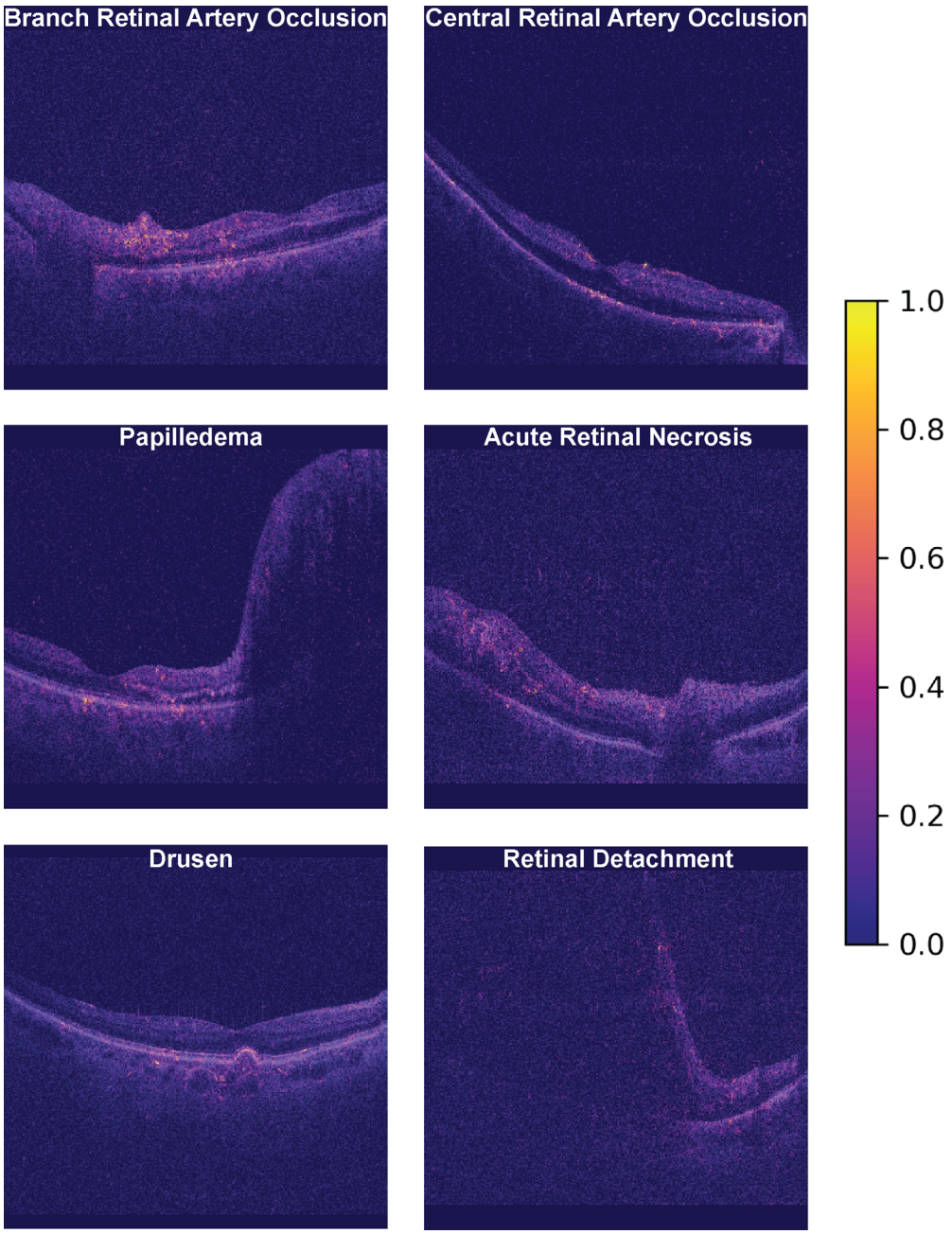

Figure 4. Deep learning heatmaps overlaid on representative OCT images with posterior segment pathologies. The heatmaps demonstrate the relative contribution of each pixel to the model's classification of these images as referable through integrated gradient attribution scores, ranging from 0 (*dark purple*) to 1 (*bright yellow*). Brighter areas represent greater contributions. Heatmaps were generated with the integrated gradients method. In general, heatmaps are more useful for images predicted to have referable pathology than images predicted to be non-referable, because in images predicted to be non-referable, the heatmap is the same for the prediction from the baseline (entirely black) image.

76% when evaluated on an independent, prospectively enrolled cohort of emergency department patients. Furthermore, our system's performance was comparable to the performance of two human experts.

Our results have important implications in the context of existing models of emergency eye care. The current standard of care instrument, the direct ophthalmoscope, has very poor performance in the hands of emergency physicians; in previous studies, only 4% to 20% of patients with clear indication for fundus examination actually received one in the ED.[39,40] Even when patients did receive a fundus

examination, a large-scale study showed that the sensitivity for acute pathology was extremely poor (0% in cases where ED providers were not already aware of an objective fundus finding identified at another facility).[40]

For screening a diverse range of posterior segment pathologies, our system compares favorably with existing approaches to improve ED eye care. For example, emergency physician review of fundus photographs obtained by a trained nurse practitioner showed a sensitivity of 46% and a specificity of 95% for relevant posterior segment pathologies.[35] Our system also compares favorably with ocular ultrasound, which has been shown to be very sensitive (94.2%) and specific (96.3%) for the diagnosis of retinal detachment but only moderately sensitive (82%) and specific (73%) for the diagnosis of papilledema.[41,42] Additionally, both fundus photography and ocular ultrasonography require specialized training in (or dedicated personnel for) image acquisition and interpretation. Key advantages of our system include the lack of specialized training for a technician/provider to acquire images, improved ability of OCT to evaluate three-dimensional structural changes in the retina and optic nerve compared with two-dimensional ocular ultrasound and fundus images, and automated classification of images, thereby reducing the requirements for clinician time, knowledge, and experience.

RobOCTNet may be a useful adjunct to assist emergency physicians with clinical decision-making. A previous study conducted by our group demonstrated that retinal OCT imaging could substantially improve emergency physicians' sensitivity for abnormalities in the posterior segment of the eye, compared with their existing standard of care examination, direct ophthalmoscopy.[25] These results complement the results from the present study because deep learning has the potential to help emergency physicians further improve their diagnostic accuracy based on OCT. Simultaneously, this prior study demonstrated that emergency physicians could make reasonable referral decisions independent of clinical decision support tools, should they disagree with the model classification.

The performance of RobOCTNet in detecting referable posterior segment pathology was comparable to the performance of two human experts performing the same task under the same condition (i.e., reviewing central foveal B-scans alone). Interestingly, expert 2 considered three (4%) central B-scan images ungradable because of quality, but expert 1 was able to grade these images, suggesting heterogeneity in clinician assessment of OCT images. Additionally, expert 2 thought clinical history was needed to make decisions about four central B-scan images, emphasizing the importance of clinical history in making referral decisions for some clinicians. These results highlight avenues for future work in deep learning-based studies, suggesting that models incorporating multimodal information (imaging and history) may be important.

The heterogeneity of diagnoses included in our external testing set is novel compared with many prior studies using deep learning/machine learning[1,2,4,5,14,28,43]; furthermore, our external testing set included patients with a broad range of severity and acuity of posterior segment pathology. Most of these studies have focused on accurately classifying images for a single condition or a few selected conditions, such as diabetic retinopathy and macular degeneration.[2,5,14,28,43] By contrast, we applied our deep learning model to accurately detect a diversity of pathologies, including pathologies not represented in the training set. For example, our model successfully identified eyes with optic nerve edema as containing referable pathology (100% sensitivity), despite having not been trained on any OCT image of optic nerve edema. We hypothesize that despite training on a limited set of pathologies, the deep learning model was able to recognize morphologies shared between different retinal and optic nerve pathologies. For example, the heatmap suggests that the model was able to recognize fluid in the retina in the context of papilledema, despite the lack of representation of papilledema in the training set. The ability of deep learning models to recognize generalizable image characteristics is a major driver behind their potential utility when applied to data different than their training data. These findings suggest future directions for deep learning in ophthalmology, which could build on the algorithm's ability to potentially approach problems in ways different than humans, address diverse clinical problems, and even reveal patterns unknown to human experts.[44,45]

Another key takeaway is the coupling of deep learning and robotics. Although many robotic systems use integrated machine learning systems to perform functions such as environmental feature detection and movement coordination,[46–49] we are not aware of integration of machine learning and robotics to achieve automated image acquisition and diagnosis in ophthalmology. This approach has the potential to help non-eye specialists, such as emergency physicians, overcome the challenge of evaluating eye conditions and facilitate effective and efficient ophthalmology referrals, because the system reduces the requirements for specialty expertise in both image acquisition and interpretation. In the context of the current state of eye care in the ED setting, where both over-triage and under-triage are unfortu-

nately commonplace,[23,24,36,50] such a system combining robotics and deep learning could substantially improve ED throughput efficiency without impairing safety.

Our study has limitations. First, our training data were composed primarily of single OCT B-scans and were supplemented with a small number of volumetric OCT images. For external testing, we provided only central foveal B-scans in the OCT volume to the model. Although this approach is consistent with prior studies,[5,51–53] there may be pathologies that are not captured in a central foveal B-scan.[54] Indeed, the only case of drusen our model missed was absent in the central foveal image supplied to RobOCTNet, and the human experts also did not detect any pathology when reviewing this image (Supplementary Fig. S2). Large training sets of volumetric data would allow for volumetric-to-volumetric training, which could improve diagnostic performance.[26] We also did not preprocess our images with segmentation. Although this segmentation-free approach has advantages such as eliminating bias from segmentation errors and minimizing preprocessing steps,[55] recent studies have shown that segmentation-based homogenization of images could improve model performance.[56,57]

Furthermore, although appropriately powered for our primary outcome of interest, our external testing cohort is a small cohort of patients enrolled in the ED of a single, large academic medical center, so it is possible that this patient population may not be generalizable to other settings, particularly to settings where there is a greater proportion of patients with milder retinal pathology.

Additionally, it is not immediately clear how our system would fit into real-world ED workflows, because this study was conducted for research purposes only and the results were not communicated to providers for real-time clinical care. The RAOCT system used in this study is an investigational device approved only for research use, and therefore we were restricted from using it to directly alter patient care. Likewise, the deep learning model has not been previously validated, and applying it directly to patient care could have negative consequences for patient safety. Future work can explore the implementation science of applying this study to real-world patient care, including immediate triage at the point of image capture, which was not explored in this study. Last, although heatmaps provide some insight into the model, they have been shown to oversimply the process by which deep learning systems work and should be interpreted with caution.[58]

Our study has several strengths. First, unlike many studies that did not validate their deep learning models in different populations,[27] we externally tested our model using an independent dataset prospectively acquired in settings and under conditions that played no role in model development. This is considered the gold-standard for model evaluation, because it has strong implications for generalizability of models to broader clinical scenarios.[27,59] Second, the diversity of pathologies represented in the external testing set, including pathologies on which the deep learning model was never trained, suggests that our results may be generalizable to a broad range of posterior eye pathologies. Third, our use of a reference standard that considered both clinical diagnosis and independent OCT review by two masked retina specialists increases the likelihood that our reference standard for model evaluation was accurate. Trustworthy labels are considered a core foundation of successful deployment of deep learning methods.[60] Our method minimizes bias from individual clinicians by using complementary clinical diagnosis and OCT information in tandem. Fourth, we used training data from a diverse range of sources, clinical settings, and OCT systems, which allowed us to create a broadly generalizable model. Importantly, most of our training data were publicly available, which obviated the need for potentially cost-prohibitive acquisition of a large amount of training data. Finally, our study design to compare the model with two human experts performing the task under comparable conditions was innovative and offered important insights about how human experts make referral decisions.

Future research could use volumetric OCT training data to further enhance the ability of deep learning models to analyze OCT data in the context of neighboring B-scans. Ensemble stacking approaches are likewise promising for integrating multiple (volumetric) OCT images to perform classification on the volume level.[61] However, these approaches will require a large amount of high-quality volumetric OCT data, which are costly to obtain and are scarce in the public domain. Increasing the availability and accessibility of such data may enable successful adoption of these approaches. Additionally, opportunities remain to evaluate the performance of our robotics-deep learning system in the context of real-world patient care, including its cost-effectiveness.

In conclusion, this study showed that combining a robotically aligned OCT system with a deep learning model, RobOCTNet, could potentially facilitate accurate detection of referable posterior segment disease among patients presenting to the emergency department. This work demonstrated the potential of coupling robotics with machine learning to improve emergency patient triage for ophthalmology referral. Future research should address key limitations of

this study, including incorporating greater amounts of volumetric OCT training data and clinical information, which may result in improved model performance.

Acknowledgments

Supported by NIH/NEI (R01-EY029302, R01-EY035534, and P30-EY005722). Ailin Song received research support from NIH/NCATS (TL1-TR002555) and Research to Prevent Blindness Medical Student Eye Research Fellowship. The sponsor or funding organization had no role in the design or conduct of this research.

Disclosure: **A. Song,** None; **J.B. Lusk,** None; **K.-M. Roh,** None; **S.T. Hsu,** None; **N.G. Valikodath,** None; **E.M. Lad,** Boehringer Ingelheim (F), provisional patent 63/162741 (P); **K.W. Muir,** None; **M.M. Engelhard,** US Patent 11308325 (P); **A.T. Limkakeng,** Roche Diagnostics, Inc. (F), Abbott Laboratories (F), Quidel Inc. (F), Brainbox Inc. (F), Forest Devices, Inc. (F), Becton Dickinson (F), SENSE Neuro Diagnostics (F), Ophirex (F); **J.A. Izatt,** Alcon, Inc. (C), Leica Microsystems (P, R); **R.P. McNabb,** Johnson & Johnson Vision (F), Leica Microsystems (P, R); **A.N. Kuo,** Johnson & Johnson Vision (F), Leica Microsystems (P, R)

References

1. Milea D, Najjar RP, Jiang Z, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. 2020;382:1687–1695.
2. 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.
3. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350.
4. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1–29.
5. 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.e9.
6. Kim JA, Yoon H, Lee D, et al. Development of a deep learning system to detect glaucoma using macular vertical optical coherence tomography scans of myopic eyes. *Sci Rep*. 2023;13(1):8040.
7. Thiéry AH, Braeu F, Tun TA, Aung T, Girard MJA. Medical application of geometric deep learning for the diagnosis of glaucoma. *Transl Vis Sci Technol*. 2023;12(2):23.
8. Li AL, Feng M, Wang Z, et al. Automated detection of posterior vitreous detachment on OCT using computer vision and deep learning algorithms. *Ophthalmol Sci*. 2023;3(2):100254.
9. Manikandan S, Raman R, Rajalakshmi R, Tamilselvi S, Surya RJ. Deep learning-based detection of diabetic macular edema using optical coherence tomography and fundus images: a meta-analysis. *Indian J Ophthalmol*. 2023;71:1783–1796.
10. Leingang O, Riedl S, Mai J, et al. Automated deep learning-based AMD detection and staging in real-world OCT datasets (PINNACLE study report 5). *Sci Rep*. 2023;13(1):19545.
11. Crincoli E, Zhao Z, Querques G, et al. Deep learning to distinguish Best vitelliform macular dystrophy (BVMD) from adult-onset vitelliform macular degeneration (AVMD). *Sci Rep*. 2022;12(1):12745.
12. Tang Y, Gao X, Wang W, et al. Automated detection of epiretinal membranes in OCT images using deep learning. *Ophthalmic Res*. 2023;66:238–246.
13. Gunasekeran DV, Wong TY. Artificial intelligence in ophthalmology in 2020: a technology on the cusp for translation and implementation. *Asia-Pac J Ophthalmol*. 2020;9:61–66.
14. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. 2019;126:552–564.
15. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. *J Big Data*. 2015;2(1):1.
16. Bureau of Labor Statistics. Ophthalmic Medical Technicians. Available at: https://www.bls.gov/oes/current/oes292057.htm. Accessed March 13, 2022.
17. Wilson FA, Stimpson JP, Wang Y. Inconsistencies exist in national estimates of eye care services utilization in the United States. *J Ophthalmol*. 2015;2015:435606.
18. Bureau UC. The U.S. Joins Other Countries With Large Aging Populations. Available at: https://www.census.gov/library/stories/2018/03/graying-america.html. Accessed March 18, 2022.
19. Committee on Public Health Approaches to Reduce Vision Impairment and Promote Eye Health, Board on Population Health and Public Health Practice, Health and Medicine Division, National Academies of Sciences, Engineering, and Medicine. *Making Eye Health a Population Health Imperative: Vision for Tomorrow*. In:

Teutsch SM, McCoy MA, Woodbury RB, Welp A, eds. Washington, DC: National Academies Press; 2016:23471.

20. Draelos M, Ortiz P, Qian R, et al. Contactless optical coherence tomography of the eyes of freestanding individuals with a robotic scanner. *Nat Biomed Eng*. 2021;5:726–736.
21. Channa R, Zafar SN, Canner JK, Haring RS, Schneider EB, Friedman DS. Epidemiology of eye-related emergency department visits. *JAMA Ophthalmol*. 2016;134:312–319.
22. Crum OM, Kilgore KP, Sharma R, et al. Etiology of papilledema in patients in the eye clinic setting. *JAMA Netw Open*. 2020;3(6):e206625.
23. Deaner JD, Amarasekera DC, Ozzello DJ, et al. Accuracy of referral and phone-triage diagnoses in an eye emergency department. *Ophthalmology*. 2021;128:471–473.
24. Nari J, Allen LH, Bursztyn LLCD. Accuracy of referral diagnosis to an emergency eye clinic. *Can J Ophthalmol*. 2017;52:283–286.
25. Song A, Roh KM, Lusk JB, et al. Robotic optical coherence tomography retinal imaging for emergency department patients: a pilot study for emergency physicians' diagnostic performance. *Ann Emerg Med*. 2023;81:501–508.
26. Yanagihara RT, Lee CS, Ting DSW, Lee AY. Methodological challenges of deep learning in optical coherence tomography for retinal diseases: a review. *Transl Vis Sci Technol*. 2020;9(2):11.
27. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167–175.
28. Srinivasan PP, Kim LA, Mettu PS, et al. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed Opt Express*. 2014;5:3568–3577.
29. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. Available at: https://www.tensorflow.org/. Accessed June 22, 2022.
30. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*; 2009:248–255.
31. Arefin R, Samad MD, Akyelken FA, Davanian A. Non-transfer deep learning of optical coherence tomography for post-hoc explanation of macular disease classification. In: *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. 2021:48–52.
32. Guan Q, Wan X, Lu H, et al. Deep convolutional neural network Inception-v3 model for differential diagnosing of lymph node in cytological images: a pilot study. *Ann Transl Med*. 2019;7(14):307.
33. Hosmer DW, Lemeshow S. Assessing the Fit of the Model. In: *Applied Logistic Regression*. Hoboken, NJ: John Wiley & Sons, Ltd; 2000:143–202.
34. Wilson CL, Leaman SM, O'Brien C, Savage D, Hart L, Jehle D. Novice emergency physician ultrasonography of optic nerve sheath diameter compared to ophthalmologist fundoscopic evaluation for papilledema. *J Am Coll Emerg Physicians Open*. 2021;2(1):e12355.
35. Bruce BB, Thulasi P, Fraser CL, et al. Diagnostic accuracy and use of non-mydriatic ocular fundus photography by emergency department physicians: phase II of the FOTO-ED Study. *Ann Emerg Med*. 2013;62(1):28–33.e1.
36. Stunkel L, Sharma RA, Mackay DD, et al. Patient harm due to diagnostic error of neuro-ophthalmologic conditions. *Ophthalmology*. 2021;128:1356–1362.
37. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
38. Sundararajan M, Taly A, Yan Q. *Axiomatic Attribution for Deep Networks*. arXiv; 2017.
39. Golombievski E, Doerrler MW, Ruland SD, McCoyd MA, Biller J. Frequency of direct funduscopy upon initial encounters for patients with headaches, altered mental status, and visual changes: a pilot study. *Front Neurol*. 2015;6:233.
40. Bruce BB, Lamirel C, Biousse V, et al. Feasibility of nonmydriatic ocular fundus photography in the emergency department: phase I of the FOTO-ED Study. *Acad Emerg Med*. 2011;18:928–933.
41. Gottlieb M, Holladay D, Peksa GD. Point-of-care ocular ultrasound for the diagnosis of retinal detachment: a systematic review and meta-analysis. *Acad Emerg Med*. 2019;26:931–939.
42. Teismann N, Lenaghan P, Nolan R, Stein J, Green A. Point-of-care ocular ultrasound to detect optic disc swelling. *Acad Emerg Med*. 2013;20:920–925.
43. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131:351–357.
44. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158–164.
45. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141):20170387.

46. Bogue R. The role of artificial intelligence in robotics. *Ind Robot Int J*. 2014;41:119–123.
47. Kunze L, Hawes N, Duckett T, Hanheide M, Krajník T. Artificial intelligence for long-term robot autonomy: a survey. *IEEE Robot Autom Lett*. 2018;3:4023–4030.
48. Mir UB, Sharma S, Kar AK, Gupta MP. Critical success factors for integrating artificial intelligence and robotics. *Digit Policy Regul Gov*. 2020;22:307–331.
49. Zhou M, Wang X, Weiss J, et al. Needle localization for robot-assisted subretinal injection based on deep learning. In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE; 2019:8727–8732.
50. Yap J, Guest S, McGhee CNJ. Characteristics and accuracy of referrals to an acute tertiary ophthalmic service in New Zealand. *Clin Experiment Ophthalmol*. 2015;43:387–389.
51. Zhang G, Fu DJ, Liefers B, et al. Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study. *Lancet Digit Health*. 2021;3(10):e665–e675.
52. Kang NY, Ra H, Lee K, Lee JH, Lee WK, Baek J. Classification of pachychoroid on optical coherence tomography using deep learning. *Graefes Arch Clin Exp Ophthalmol*. 2021;259:1803–1809.
53. 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.
54. McNabb RP, Grewal DS, Mehta R, et al. Wide field of view swept-source optical coherence tomography for peripheral retinal disease. *Br J Ophthalmol*. 2016;100:1377–1382.
55. Thompson AC, Jammal AA, Berchuck SI, Mariottoni EB, Medeiros FA. Assessment of a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans. *JAMA Ophthalmol*. 2020;138:333–339.
56. Russakoff DB, Lamin A, Oakley JD, Dubis AM, Sivaprasad S. Deep learning for prediction of AMD progression: a pilot study. *Invest Ophthalmol Vis Sci*. 2019;60:712–722.
57. Russakoff DB, Mannil SS, Oakley JD, et al. A 3D deep learning system for detecting referable glaucoma using full OCT macular cube scans. *Transl Vis Sci Technol*. 2020;9(2):12.
58. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745–e750.
59. Cheung CY, Tang F, Ting DSW, Tan GSW, Wong TY. Artificial intelligence in diabetic eye disease screening. *Asia-Pac J Ophthalmol*. 2019;8:158–164.
60. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Med Image Anal*. 2020;65:101759.
61. Wolpert DH. Stacked generalization. *Neural Netw*. 1992;5:241–259.