tvst

Artificial Intelligence

Deep Learning Algorithm Detects Presence of Disorganization of Retinal Inner Layers (DRIL)–An Early Imaging Biomarker in Diabetic Retinopathy

Rupesh Singh[1,*], Srinidhi Singuri[2,*], Julia Batoki[1], Kimberly Lin[1], Shiming Luo[2], Dilara Hatipoglu[3], Bela Anand-Apte[1], and Alex Yuan[1]

[1] Cole Eye Institute, Cleveland Clinic, Cleveland, OH, USA
[2] Cleveland Clinic Lerner College of Medicine, Cleveland, OH, USA
[3] Case Western Reserve University, Cleveland, OH, USA

Correspondence: Rupesh Singh, 9500 Euclid Ave i32, Cleveland, OH 44195, USA. e-mail: rupesh_singh@live.in

Received: January 6, 2023
Accepted: June 8, 2023
Published: July 6, 2023

Keywords: disorganization of retinal inner layers (DRIL); diabetic retinopathy (DR); deep learning (DL); convolution neural network (CNN); artificial intelligence

Citation: Singh R, Singuri S, Batoki J, Lin K, Luo S, Hatipoglu D, Anand-Apte B, Yuan A. Deep learning algorithm detects presence of disorganization of retinal inner layers (DRIL)–An early imaging biomarker in diabetic retinopathy. Transl Vis Sci Technol. 2023;12(7):6, https://doi.org/10.1167/tvst.12.7.6

Purpose: To develop and train a deep learning-based algorithm for detecting disorganization of retinal inner layers (DRIL) on optical coherence tomography (OCT) to screen a cohort of patients with diabetic retinopathy (DR).

Methods: In this cross-sectional study, subjects over age 18, with ICD-9/10 diagnoses of type 2 diabetes with and without retinopathy and Cirrus HD-OCT imaging performed between January 2009 to September 2019 were included in this study. After inclusion and exclusion criteria were applied, a final total of 664 patients (5992 B-scans from 1201 eyes) were included for analysis. Five-line horizontal raster scans from Cirrus HD-OCT were obtained from the shared electronic health record. Two trained graders evaluated scans for presence of DRIL. A third physician grader arbitrated any disagreements. Of 5992 B-scans analyzed, 1397 scans (∼30%) demonstrated presence of DRIL. Graded scans were used to label training data for the convolution neural network (CNN) development and training.

Results: On a single CPU system, the best performing CNN training took ∼35 mins. Labeled data were divided 90:10 for internal training/validation and external testing purpose. With this training, our deep learning network was able to predict the presence of DRIL in new OCT scans with a high accuracy of 88.3%, specificity of 90.0%, sensitivity of 82.9%, and Matthews correlation coefficient of 0.7.

Conclusions: The present study demonstrates that a deep learning-based OCT classification algorithm can be used for rapid automated identification of DRIL. This developed tool can assist in screening for DRIL in both research and clinical decision-making settings.

Translational Relevance: A deep learning algorithm can detect disorganization of retinal inner layers in OCT scans.

Introduction

Diabetic retinopathy (DR) is a common sight-threatening complication of diabetes and represents a leading cause of blindness among working-aged adults worldwide.[1] Recent population-based studies estimate a global prevalence of 103.12 million adults with DR.[1] Of that population, a predicted 28.54 million adults experience sight-threatening disease.[1]

Although current therapies for DR and diabetic macular edema (DME) are effective, early detection of disease and prompt management facilitate optimal treatment outcomes.[2] Use of accurate screening tools that refer the correct population of patients to retina specialists is key. Optical coherence tomography (OCT) is one imaging modality that is routinely used in clinical practice for screening patients for diabetic retinopathy. Although certain OCT imaging biomarkers are established (i.e., intraretinal fluid for DME), the

clinical significance of other biomarkers, such as disorganization of retinal inner layers (DRIL), remains to be determined.[3,4] Presence of DRIL has been shown to be associated with reduced visual acuity,[5–7] reduced retinal function,[8] ellipsoid zone disruption,[9] and thinning of retinal nerve fiber layer.[9] However, DRIL can be challenging to detect, because OCT changes may be subtle, especially in early or mild cases of DR.[7]

Recent advances in artificial intelligence have shown great potential in rapid, automated, and accurate medical decision-making in ophthalmology.[10,11] Deep learning (DL)—a subset of machine learning and artificial intelligence—is a convolution neural network (CNN)–based learning algorithm that can be applied to the interpretation and classification of clinical imaging. Trained and optimized, CNNs have the capability to identify unique imaging features and classify images with high accuracy.[10,11] Recently, OCT-based DL algorithms have been used to detect central serous chorioretinopathy,[12] ellipsoid zone defects,[13] and intraretinal fluid.[14]

Accurate and automated detection of DRIL by DL-based algorithms represents one strategy for standardizing interpretation of DRIL. In this study, we modify pretrained CNNs to create a neural network best suited to identifying the imaging biomarker DRIL, in OCT imaging from a cohort of diabetic patients.

Methods

Subject Selection

Study subjects were selected for inclusion in a cross-sectional study approved by the Cleveland Clinic Foundation Institutional Review Board. Research adhered to the tenets of the Declaration of Helsinki and complied with Health Insurance Portability and Accountability Act privacy and security regulations.

Table 1 represents the demographic of included subjects. Inclusion criteria for subjects included age over 18, ICD-9/10 diagnoses of type 2 diabetes with and without retinopathy, and Zeiss Cirrus HD-OCT

Table 1. Sample Characteristics of Entire Study Cohort and by DRIL Status

		DRIL Status*	
Sample Characteristics Count (%)	Total $n = 1201$	DRIL $n = 410$ (34.1%)	No DRIL $n = 791$ (65.9%)
DR Severity			
No DR	689 (57.4%)	94 (22.9%)	595 (75.2%)
Mild NPDR	199 (16.6%)	88 (21.5%)	111 (14.0%)
Moderate NPDR	88 (7.3%)	50 (12.2%)	38 (4.8%)
Severe NPDR	54 (4.5%)	40 (9.6%)	14 (1.8%)
PDR	171 (14.2%)	138 (33.7%)	33 (4.2%)
Cysts			
No cyst present	918 (76.4%)	189 (46.1%)	729 (92.2%)
Some form of cyst present	283 (23.6%)	221 (53.9%)	62 (7.8%)
Demographics			
Sex			
Female	638 (53.1%)	206 (50.2%)	432 (54.6%)
Male	563 (46.9%)	204 (49.8%)	359 (45.4%)
Age (yr),† mean [SD]	69.0 [11.0]	67.3[11.0]	70.0[11.6]
Ethnicity			
White	707 (58.9%)	226 (55.1%)	481 (60.8%)
Asian	25 (2.1%)	10 (2.4%)	15 (1.9%)
Black	428 (35.6%)	160 (39.0%)	268 (33.9%)
Other‡	41 (3.4%)	14 (3.4%)	27 (3.4%)

Single column descriptive statistics presented for the total study sample and stratified by DRIL status. Categorical variables reported with frequencies and percentage.

†Continuous variables reported with mean and standard deviation (SD).

*Percentages are of n = number of eyes.

‡Ethnicity of other includes multiracial/multicultural, declined, unavailable, American Indian, and Alaskan native. Total of 1201 eyes from 664 patients were used with 537 patients had both eyes imaging.

imaging performed at a tertiary eye care center between January 2009 to September 2019. Exclusion criteria included presence of any retinal dystrophy, macular or lamellar holes, retinal vein or artery occlusion, wet age-related macular degeneration, vitreomacular traction altering the foveal contour, or poor image quality (i.e., severe motion artifact, or dim view). Cysts (of any size) were not excluded from the study. If the cyst disrupted the layers in such a way that disrupted the transition line between two zones (i.e., IPL/INL) we considered this DRIL. However, if the cyst simply displaced the transition line and all transition lines for each layer were still discernible, we would not consider that DRIL. Epiretinal membrane (ERM), because of their epiretinal nature were not excluded. If the ERM was associated with significant vitreomacular traction that altered the foveal contour in a significant way such that the zones were no longer clearly visible, we excluded the scan.

An automated query of the electronic health record was used to generate a list of eligible study subjects. The list was ordered by medical record number—a randomly generated eight-digit number unique to every subject. After inclusion and exclusion criteria were applied, the first 664 subjects—a total of 5992 B-scans—were included for analysis. Total of 1201 eyes were used with 537 patients had both eyes imaging. We excluded some eyes because of poor image quality. Moreover, 127 patients had only one eye examined or imaged. Whenever both eyes of the patients were available, both eyes were included in the study.

Data Collection

Cirrus HD-OCT images were acquired from the electronic health record. For this study, DRIL was defined as any disruption of the inner nuclear layer (INL), outer plexiform layer (OPL), and the ganglion cell layer-inner plexiform layer (GCL-IPL) in the central 1000 um of the fovea, as defined by Sun et al., 2014.[3] Individual OCT scans (high definition, 5-line, horizontal raster scans) were evaluated for presence of DRIL by masked graders trained by a retina specialist. A third masked grader (retina specialist) resolved disagreements. Cohen's kappa was calculated to determine agreement between graders. Final grading of scans was used to label the images used for subsequent CNN model development. Images were downloaded to secure database and deidentified before grading.

Model Development and Testing

To develop the DL-based algorithm, four pre-existing networks were modified for binary classification of images: "DRIL" or "no DRIL." These networks were Alexnet,[15] GoogLeNet,[16] InceptionResNetV2,[17] and NasNetLarge.[18] Networks were taught to detect DRIL by three distinctive steps: network training, network validation, and external testing.

Graded OCT images were randomly split into three subsets corresponding to the steps of model development. Of all labeled images, 72% were selected for network training, 18% were selected for network validation, and the remaining 10% were reserved for external testing.

During network training, labeled images and DRIL designation were fed to networks to teach algorithms to recognize scans of "DRIL" and "no DRIL." Interspersed within the network training process were network validation steps. During network validation, algorithms were given a labeled image with the DRIL designation initially withheld. The algorithm predicted the presence of DRIL based on what it had learned thus far in the training process, and told whether its determination was correct or incorrect. Based on this feedback, the network modified itself to improve. In the last step of model development, trained and validated networks were used to test an external data set. In this step, images were fed to networks to test their ability to accurately classify scans by the presence of DRIL. Network output was compared against manual grading of the same scans.

All CNNs were trained and tested with and without initial weight freeze. Further, data augmentation[19] in conjunction with initial weight freeze was used for retraining GoogLeNet. Data augmentation used slightly modified version of images; image rotation (three images with up to 30° rotation) and y-axis reflection of images were fed to the networks for training.[19] The training of all four modified CNNs was performed using MATLAB (MathWorks Inc., Natick, MA, USA) on a single computer (8-core CPU. Windows 10; Microsoft, Redmond, WA, USA) with parallel processing using eight workers.

Finally, gradient-weighted class activation mapping (Grad-CAM) was performed on randomly selected OCT scans with GoogLeNet. Grad-CAM maps visually demonstrate where in the image the learned decision-making is occurring.[20]

Statistical Analysis

The accuracy and area under the curve (AUC) from both, internal validation and external testing steps were obtained for all four CNNs trained with and without initial weight freeze. GoogLeNet trained with initial weight freeze and data augmentation was analyzed for accuracy and AUC for internal validation

and external testing. A rigorous set of statistical measures were calculated for all three modified GoogLeNet algorithms: accuracy, AUC, error rate, false-positive rate, false-negative rate, specificity, sensitivity, precision, F-1 score and Matthew correlation coefficient (MCC).[21] Results were plotted on receiver operating characteristic (ROC) curves and presented on confusion matrices and tables.

Results

A total of 5992 OCT B-scans, from a cohort of diabetic patients, were manually evaluated for presence of DRIL by trained graders. Cohen's kappa was greater than 0.85, indicating near-perfect agreement between graders.

Graded OCT images were split into three subsets for development and training of CNNs. A subset of 72% of all graded OCT images (n = 4328) were used solely for network training. Another subset of 18% of all graded images (n = 1082) were used for internal network validation steps. The remaining 10% of images (n = 600) were used for external testing of networks.

Four different pretrained CNNs—Alexnet,[15] GoogLeNet,[16] InceptionResNetV2,[17] and NasNet-Large[18]—were trained with and without initial weight freeze and tested for their ability to correctly classify and predict presence of DRIL. Network accuracy and AUC for internal validation and external testing steps were compared for each CNN (Table 2). GoogLeNet and InceptionResNetV2 with initial weights freeze resulted in high accuracy (85.8% and 88.7%) and AUC (0.91 and 0.90), respectively.

The typical training time of CNNs depended on size and complexity. On a single CPU, training with initial weight freeze using Alexnet was ~32 minutes, GoogLeNet was ~54 minutes, InceptionResNetV2 was ~406 minutes, and NasNetLarge was ~2292 minutes. Based on accuracy, AUC, training time, sensitivity, and specificity, GoogLeNet was selected for further analysis.

GoogLeNet was retrained using initial weight freeze and data augmentation. ROC curves of the three modified GoogLeNet algorithms—(1) without initial weight freeze, (2) with initial weight freeze, and (3) with initial weight freeze and data augmentation—were produced and AUC for each were calculated (Fig. 1). AUC of ROC curves was highest for GoogLeNet modified with initial weight freeze and data augmentation (AUC = 0.93), compared to GoogLeNet modifications without initial weight freeze (AUC = 0.92) and with initial weight freeze (AUC = 0.91).

Based on DRIL determinations made by networks, performance of the three GoogLeNet modified networks was further analyzed by eight parametric/statistical tests. Results were presented in confusion matrices/contingency tables (Fig. 2) and summarized in Tables 3 and 4.[22] External testing of GoogLeNet modified with initial weight freeze and data augmentation resulted in high accuracy (88.3%), high specificity and precision (90.0% and 82.9%, respectively), low error rate and false-positive rate (11.0% and 10.0%, respectively), an F1-score of 76.9%, and an MCC score of 0.7.

Representative examples of CNN's ability to predict the presence of DRIL are presented in Figure 3. DRIL-classified OCT images are accompanied by the probability of the prediction as determined by the deep-learning algorithm (Fig. 3).

Last, data from Grad-CAM were presented as a heat map showing the focus of learned networks during decision making (Fig. 4).[20] Our CNN was able to detect DRIL in diabetic patients, irrespective of DR severity or the presence of DME.

Discussion

In this article, we developed and trained a deep learning algorithm for detecting the OCT-imaging

Table 2. Accuracy and AUC of Four CNNs Tested During Internal Validation and External Testing Steps With and Without Initial Weight Freeze

	Without Weight Freeze				With Weight Freeze			
	Internal		External		Internal		External	
CNN Name	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Alexnet	88.7%	0.93	88.0%	0.90	90.5%	0.94	85.8%	0.92
GoogLeNet	81.9%	0.95	79.1%	0.92	89.9%	0.95	85.8%	0.91
InceptionResNetV2	89.5%	0.94	85.8%	0.89	89.2%	0.93	88.7%	0.90
NasNetLarge	88.5%	0.92	86.3%	0.92	88.5%	0.93	86.5%	0.89

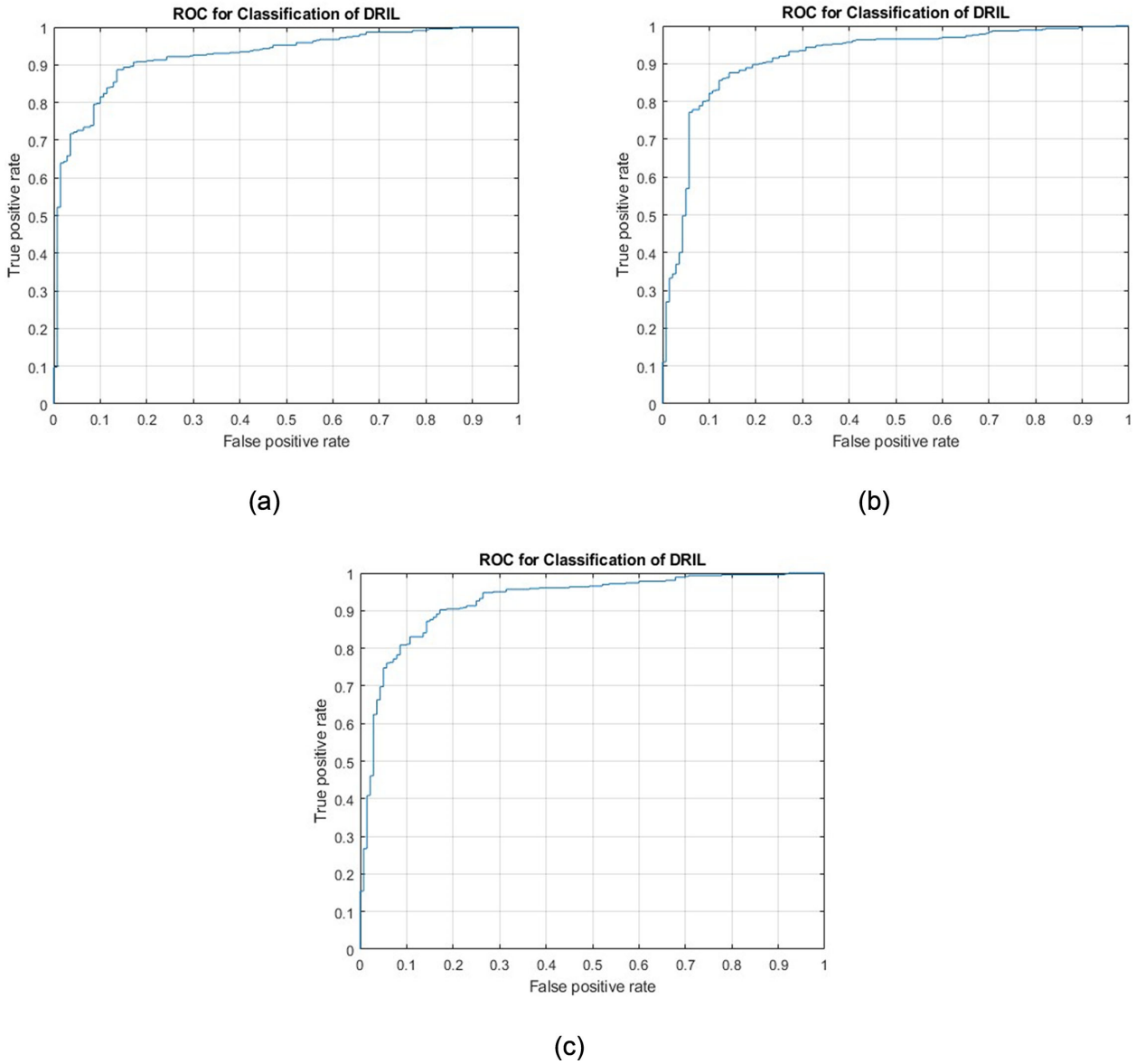

Figure 1. The receiver operating characteristic curves for modified GoogLeNet (**a**) without initial weight freeze, AUC = 0.92, (**b**) with initial weight freeze, AUC = 0.91, and (**c**) with image data augmentation and initial weight freeze, AUC = 0.93.

biomarker, DRIL in a cohort of diabetic subjects with and without retinopathy. Using a transfer-learning protocol, we modified pre-existing neural networks—originally designed and used for other purposes—to detect DRIL on OCT.[23–25]

Initial selection of pretrained CNNs in the present study was made to show the feasibility of achieving a clinically meaningful decision-making algorithm using transfer learning. Use of this method allowed for efficient training of CNNs with a relatively small data set and reduced training time.[25] In the present article, retraining required a comparatively small data set (e.g., few thousand images) compared with the larger data set that would be needed for new and un-trained CNNs (e.g., millions of images).[26]

Multiclass pretrained CNNs were modified for binary classification of DRIL (i.e., "yes – DRIL present," "no – DRIL not present"). It is important to note that typically, for binary classification, a relatively small training data set (when compared to multiclass CNNs) yields a robust CNN.[26,27] Development of the algorithm was focused on achieving high accuracy (i.e., true positive and true negative) and minimal errors (i.e., false positive and false negative) in classification of images with DRIL. To improve accuracy, algorithms were further modified by testing with and

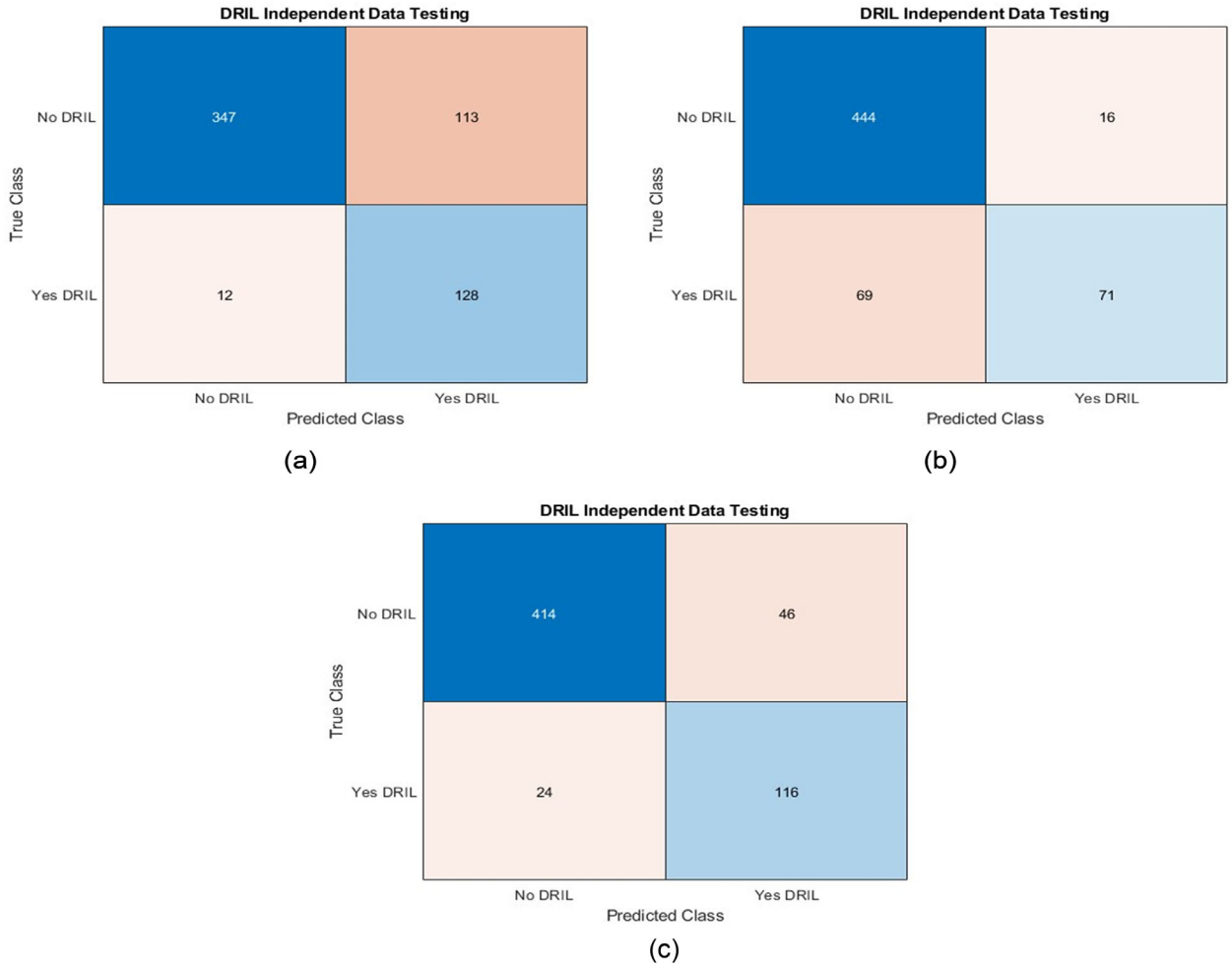

Figure 2. Confusion matrices summarizing results from external testing of modified GoogLeNet (**a**) without initial weight freeze, (**b**) with initial weight freeze, and (**c**) with image data augmentation and initial weight freeze.

Table 3. Result Parameters for Modified GoogLeNet Algorithms for Classification of Presence of DRIL From External Testing

Modified GoogLeNet	Without Weight Freeze	With Weight Freeze	With Weight Freeze and Data Augmentation
Total images classified	600	600	600
Actual yes	140	140	140
Actual no	460	460	460
True positive	128	71	116
True negative	347	444	414
False positive	113	16	46
False negative	69	69	24

without initial weight freezing. The initial layers weight freeze is a technique to improve efficiency of the CNN, especially in transfer learning protocols. When we retrain a pretrained CNN, the weights between the neural network layers are updated according to the new task. For first few layers, freezing the weights of pretrained neural network from updating help the CNN to be retrained more quickly, more efficiently, and with less data required for achieving similar results without initial weight freeze.[28,29] The freezing of initial

Table 4. Classification Parameters for Modified GoogLeNet Algorithms From External Testing

Parameter	Definition *100	Without Weight Freeze	With Weight Freeze	With Weight Freeze and Data Augmentation
Accuracy	TP + TN/Total	79.2%	85.8%	88.3%
Error rate (misclassification rate)	FP + FN/Total	30.3%	14.2%	11.7%
Sensitivity or recall (true positive rate)	TP/Actual yes	91.4%	50.0%	82.9%
False positive rate	FP/Actual no	24.5%	3.5%	10.0%
Specificity (true negative rate)	TN/Actual no	75.4%	96.5%	90.0%
Precision	TP/Predicted yes	53.11%	81.6%	71.6%
F1-score	TP/(TP + 0.5 (FP + FN))	58.44%	62.6%	76.8%
MCC (Ranges from −1 to 1, with 1 being the best)	$TP \times TN - FP \times FN/\sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}$	0.4	0.6	0.7

TP, true positive; TN, true negative; FP, false positive; FN, false negative.

weights of pretrained networks results in faster training and validation. In the present article, an initial 10-layer weight freeze while training resulted in higher accuracy and AUC of CNNs.

We compared the four modified CNNs to determine that GoogLeNet was the best choice for further algorithm optimization. The selection of GoogLeNet (for data augmentation training and testing) over other three models was made by comparing crucial initial training and testing parameters (with both internal validation data and external testing data). The training time of GoogLeNet was significantly lower whereas the performance parameters were comparable to larger CNNs (e.g., InceptionResNetV2 and NasNetLarge).

Our modified GoogLeNet was retrained with initial weight freeze[30] and data augmentation,[19] as both of these techniques have been shown to improve CNN efficiency of image classification. Data augmentation refers to slight alterations of images (rotation and inversion) fed to networks to increase the training sample. The augmented data set had significantly improved the accuracy, F-1 score, and MCC score compared with other arrangements of training CNNs (e.g., without and with initial weight freeze).

Statistical tests were performed to compare efficacies of the modified GoogLeNet algorithms. A robust CNN for detection and classification of diseases has strong attributes in terms of high accuracy, low error rates, and high F1 and MCC scores.[12,31] The accuracy was obtained for independent data classification. However, accuracy alone can be misleading and biased when assessing binary classification CNN performance. Accuracy of networks must be interpreted in the context of other statistical measures such as low error rate or misclassification rate as a measure of the CNN's ability to avoid wrong classifications.

Other parameters measured were F-1 scores and MCC. The F-1 score is a harmonic mean of precision and sensitivity and has been established in the literature as a metric—alongside accuracy—to evaluate machine learning. A higher value of F-1 score indicates a healthy balance between sensitivity and precision. However, more recently, MCC has been shown to be a more reliable and robust measure of CNN performance compared to both, accuracy and F-1 score.[21] Unlike other parameters mentioned in Table 4, the MCC ranges from −1 to 1 (with −1 being the worst and 1 being the best).[21] CNN accuracy, F-1 score, and MCC scores improved with initial weight freeze and data augmentation when compared to the original modified pretrained network (e.g., without weight freeze) for binary classification (Table 4).

Finally, Grad-CAM was used to determine the learned area of decision making within images of the neural network. Grad-CAM heat maps demonstrate that the focus of the learned network is in the central fovea when determining presence of DRIL. This finding is expected, and encouraging, because our parameters for grading DRIL focused on the central 1000 µm of the retina. The Grad-CAM images confirm that the neural network has been trained to focus on central retina when identifying DRIL on OCT scans. All parameters (Table 4) and Grad-CAM images (Fig. 4) indicate production of an accurate and robust end-to-end deep learning algorithm for classification of DRIL in diabetic patients.

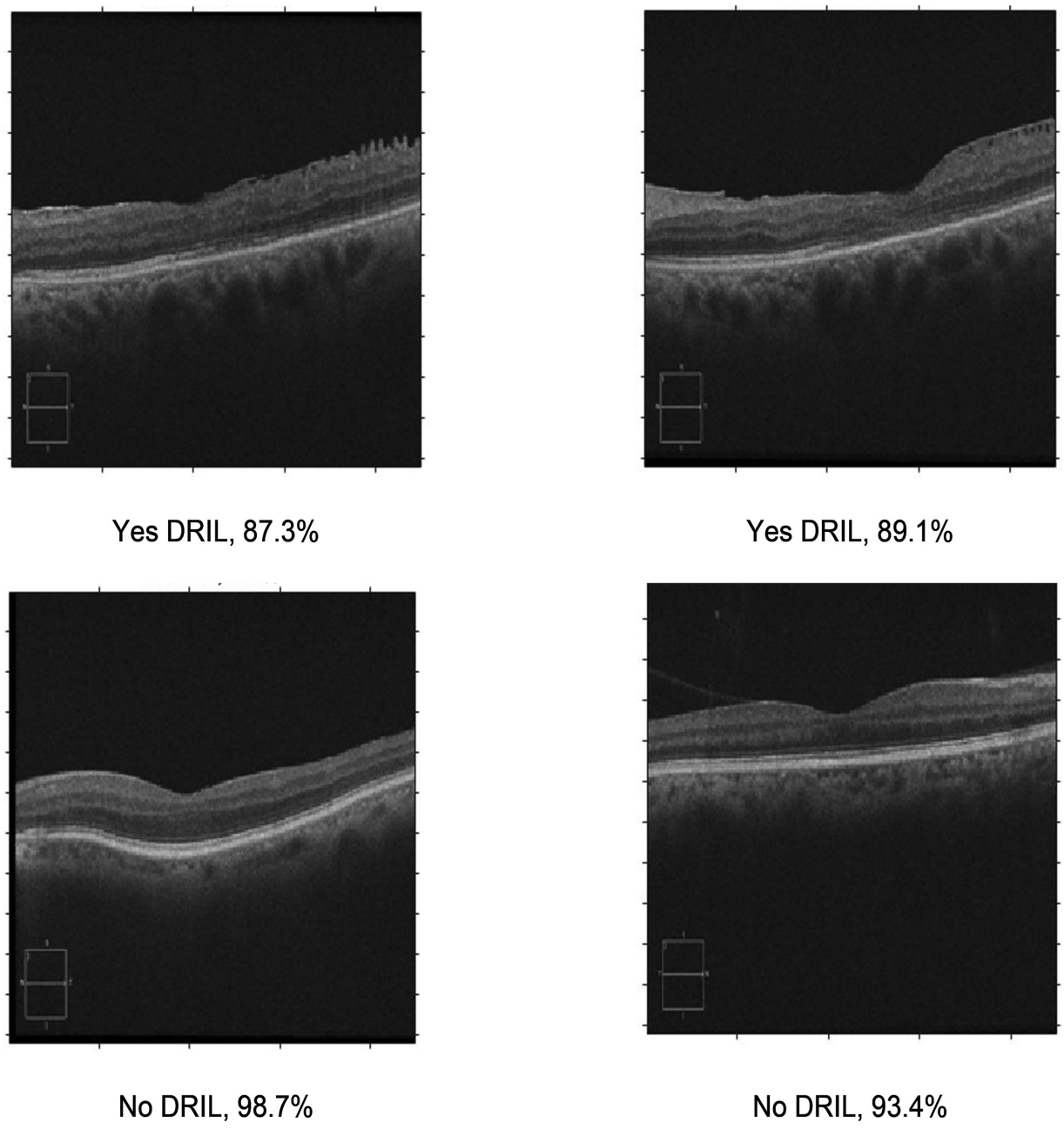

Figure 3. Representative images demonstrating the CNN's ability to detect the presence of DRIL using the modified GoogLeNet with initial weight freeze and image data augmentation.

In some of scans, foveal contour was altered by the ERM without vitreomacular traction. Clinically, these alterations in the foveal contour by ERM is sometimes of little to no visual consequence whereas vitreomacular traction with large disruptions in the foveal contour is most often associated with blurred vision. We were able to reliably grade scans with mild/moderate foveal disruption caused by ERM alone for the presence of DRIL (such as the example in Fig. 3). Only alteration of the foveal contour by vitreomacular traction was excluded from the study. Presence of ERM and/or foveal contour alteration was acceptable, and if all the layers were visible, the scan was graded and included. Thus we did not exclude scans based on ERM alone.

Future directions include developing a deep learning algorithm for multi-class classification of presence and severity of DRIL, DR, and DME. To demonstrate the ability of the deep learning algorithm to classify DRIL without confounding effects from ERM or cysts, we selected an additional cohort of patients with and without cysts and ERM for analysis (Supplement 1).

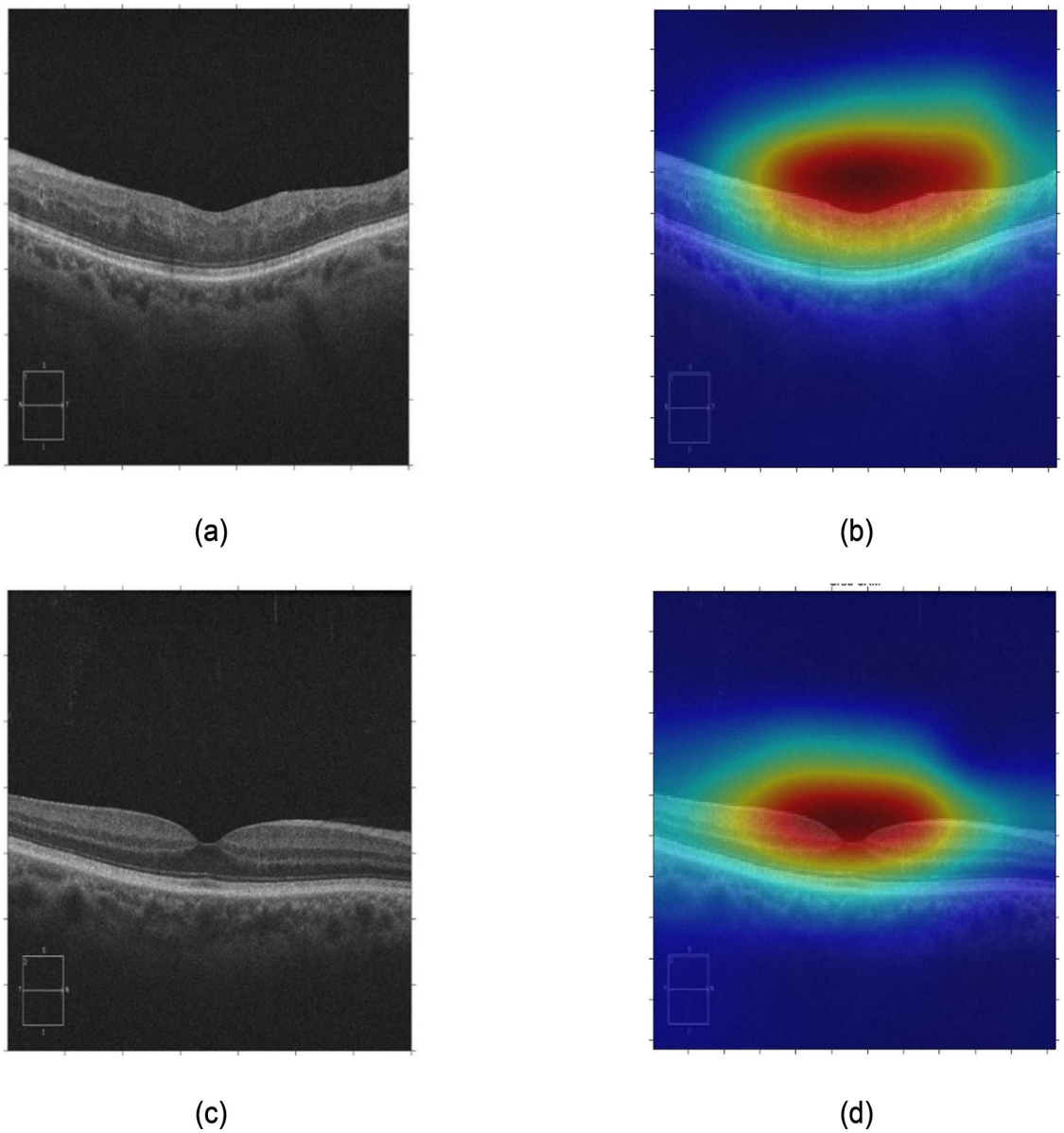

Figure 4. Representative OCT B-scans (left) with corresponding (right) grad-CAM maps of modified GoogLeNet predicted scans. The top two rows (**a**, original; **b**, prediction) are for Yes—DRIL prediction, and the bottom two rows (**c**, original; **d**, prediction) are for No—DRIL predictions.

Conclusion

Development of a DL algorithm for the recognition of DRIL in patients with DR presents unique opportunities. DRIL has been associated with worse visual prognosis, and thus its identification in patients—perhaps as part of OCT screening programs—may help identify those who would benefit from early intervention and referral to retina specialists. Furthermore, its identification at baseline in clinical trials may be important to help stratify patients based on prognosis. Manual grading of DRIL is tedious and time consuming, and automation may help allow large-scale studies to explore associations between development of DRIL and response to treatment.

Acknowledgments

Supported by RO1 grant EY026181, Unrestricted Grant to Cole Eye Institute, Research to Prevent Blindness, Research to Prevent Blindness Allergan Medical

Student Fellowship, Research to Prevent Blindness. The funding organization had no role in the design or conduct of this research.

Presented at the ARVO Imaging in the Eye Conference Abstract, July 2020.

Disclosure: **R. Singh**, None; **S. Singuri**, None; **J. Batoki**, None; **K. Lin**, None; **S. Luo**, None; **D. Hatipoglu**, None; **B. Anand-Apte**, None; **A. Yuan**, None

[*] RS and SS contributed equally to this work.

References

1. Teo ZL, Tham YC, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*. 2021;128:1580–1591.
2. Wong TY, Sun J, Kawasaki R, et al. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*. 2018;125:1608–1622.
3. Sun JK, Lin MM, Lammer J, et al. Disorganization of the retinal inner layers as a predictor of visual acuity in eyes with center-involved diabetic macular edema. *JAMA Ophthalmol*. 2014;132:1309–1316.
4. Das R, Spence G, Hogg RE, Stevenson M, Chakravarthy U. Disorganization of inner retina and outer retinal morphology in diabetic macular edema. *JAMA Ophthalmol*. 2018;136:202–208.
5. Babiuch AS, Han M, Conti FF, Wai K, Silva FQ, Singh RP. Association of disorganization of retinal inner layers with visual acuity response to anti–vascular endothelial growth factor therapy for macular edema secondary to retinal vein occlusion. *JAMA Ophthalmol*. 2019;137:38–46.
6. Radwan SH, Soliman AZ, Tokarev J, Zhang L, van Kuijk FJ, Koozekanani DD. Association of disorganization of retinal inner layers with vision after resolution of center-involved diabetic macular edema. *JAMA Ophthalmol*. 2015;133:820–825.
7. Zur D, Iglicki M, Feldinger L, et al. Disorganization of retinal inner layers as a biomarker for idiopathic epiretinal membrane after macular surgery—the DREAM Study. *Am J Ophthalmol*. 2018;196:129–135.
8. Joltikov KA, Sesi CA, de Castro VM, et al. Disorganization of retinal inner layers (DRIL) and neuroretinal dysfunction in early diabetic retinopathy. *Invest Ophthalmol Vis Sci*. 2018;59:5481–5486.
9. Nadri G, Saxena S, Stefanickova J, et al. Disorganization of retinal inner layers correlates with ellipsoid zone disruption and retinal nerve fiber layer thinning in diabetic retinopathy. *J Diabetes Complications*. 2019;33:550–553.
10. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: the technical and clinical considerations. *Progr Retinal Eye Res*. 2019;72:100759.
11. Gerendas BS, Bogunović H, Schmidt-Erfurth U. Deep learning–based automated optical coherence tomography segmentation in clinical routine: getting closer. *JAMA Ophthalmol*. 2021;139:973–974.
12. Yoon J, Han J, Park JI, et al. Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. *Sci Rep*. 2020;10:18852.
13. Loo J, Clemons TE, Chew EY, Friedlander M, Jaffe GJ, Farsiu S. Beyond performance metrics: automatic deep learning retinal OCT analysis reproduces clinical trial outcome. *Ophthalmology*. 2020;127:793–801.
14. Ehlers JP, Clark J, Uchida A, et al. Longitudinal higher-order OCT assessment of quantitative fluid dynamics and the total retinal fluid index in neovascular AMD. *Transl Vis Sci Technol*. 2021;10:29–29.
15. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Red Hook, NY: Curran Associates, Inc; 2012:84–90.
16. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Red Hook, NY: Curran Associates, Inc; 2015:1–9.
17. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. Washington, DC: AAAI Press; 2017.
18. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Red Hook, NY: Curran Associates, Inc; 2018:8697–8710.
19. Bar-David D, Bar-David L, Soudry S, Fischer A. Impact of data augmentation on retinal OCT image segmentation for diabetic macular edema analysis. In: Fu H, Garvin MK, MacGillivray T, Xu Y, Zheng Y, eds. *Ophthalmic Medical Image*

Analysis. Berlin: Springer International Publishing; 2021:148–158.

20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017:618–626.
21. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21:6.
22. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*. 2019;91:216–231.
23. Mesnil G, Dauphin Y, Glorot X, et al. Unsupervised and transfer learning challenge: a deep learning approach. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. Philadelphia: PMLR. 2012:97–110.
24. Bengio Y. Deep Learning of Representations for Unsupervised and Transfer Learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*. Philadelphia: PMLR. 2012:17–36.
25. Zhao W. Research on the deep learning of the small sample data based on transfer learning. *AIP Conference Proc*. 2017;1864:020018.
26. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88.
27. Olson M, Wyner AJ, Berk R. Modern neural networks generalize on small data sets. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc; 2018:31.
28. Putzu L, Piras L, Giacinto G. Convolutional neural networks for relevance feedback in content based image retrieval. *Multimedia Tools Appl*. 2020;79:26995–27021.
29. Venerito V, Angelini O, Cazzato G, et al. A convolutional neural network with transfer learning for automatic discrimination between low and high-grade synovitis: a pilot study. *Intern Emerg Med*. 2021;16:1457–1465.
30. Soekhoe D, van der Putten P, Plaat A. On the impact of data set size in transfer learning using deep neural networks. In: *Advances in Intelligent Data Analysis XV*. Boström H, Knobbe A, Soares C, Papapetrou P, eds. Berlin: Springer International Publishing; 2016.
31. Gjoreski M, Gams MŽ, Luštrek M, Genc P, Garbas JU, Hassan T. Machine learning and end-to-end deep learning for monitoring driver distractions from physiological and visual signals. *IEEE Access*. 2020;8:70590–70603.