Non-transfer Deep Learning of Optical Coherence Tomography for Post-hoc Explanation of Macular Disease Classification

Raisul Arefin
Dept. of Computer Science
Auburn University
Auburn, AB USA
ran0013@auburn.edu

Manar D. Samad
Dept. of Computer Science
Tennessee State University
Nashville, TN USA
msamad@tnstate.edu

Furkan A. Akyelken
Dept. of Computer Science
Tennessee State University
Nashville, TN, USA
fakyelke@tnstate.edu

Arash Davanian
Vanderbilt Eye Institute
Vanderbilt University Medical Center
Nashville, TN, USA
arash.davanian@vumc.org

2021 IEEE 9th International Conference on Healthcare Informatics (ICHI) | 978-1-6654-0132-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICHI52183.2021.00020

Abstract—**Deep transfer learning is widely used for medical image classification by leveraging models that are pretrained by natural images. This choice may introduce unnecessary model complexity that can limit explanations of such model outcomes in clinical practice. To investigate this hypothesis, we develop a configurable deep convolutional neural network (CNN) to classify four macular disease types using retinal optical coherence tomography (OCT) images. Our proposed non-transfer deep CNN model (acc: 97.9%) outperforms existing transfer learning models such as ResNet-50 (acc: 89.0%), ResNet-101 (acc: 96.7%), VGG-19 (acc: 93.3%), and Inception-V3 (acc: 95.8%) in the same retinal OCT image classification task. Our post-hoc analysis of the model extracted image features reveals that only eight out of 256 CNN filter kernels are active at the final convolutional layer. The convolutional responses of these eight selective filters yield image features that efficiently separate four macular disease classes even when projected onto two-dimensional principal component space. Our findings suggest that a large portion of deep learning parameters and computations are redundant for retinal OCT image classification, which intensifies when using transfer learning. Additionally, we provide clinical interpretations of our misclassified test images identifying manifest artifacts, shadowing of useful texture, false texture representing fluids, and other confounding factors. These clinical explanations along with model optimization via filter selection can improve the classification accuracy, computational costs, and explainability of deep model outcomes.**

Index Terms—**Convolutional neural network, Macular disease, Ophthalmology, Explainable AI, Feature extraction, Filter kernels, Medical imaging**

I. INTRODUCTION

Optical coherence tomography (OCT) can capture a cross-sectional view of the retina and optic nerve at micrometer resolutions. This imaging modality has become well-recognized for the diagnosis and management of common retinal conditions such as age-related macular degeneration (AMD) and diabetic macular edema (DME). Both disease processes can affect the macula, which is the anatomic structure responsible for detailed visual acuity. A clinician's visual assessment of OCT images may still miss subtle but important differential features for the diagnosis and grading of disease conditions. To assist diagnostic imaging, recent advances in computer vision algorithms can learn an image at multiple hierarchical layers of representations, which is known as deep learning. Several variants of deep convolutional neural network (CNN) have achieved high accuracy in the classification of retinal OCT images [1]–[3].

The image classification literature frequently uses pretrained deep learning modules, known as transfer learning, to maximize classification accuracy. The sole focus on classification accuracy results in millions of parameters within deep layer-wise model, which eventually turns into an opaque (hard-to-interpret) black-box [4]. Notably, clinical experts expect to see far more insights into their data beyond classification accuracy to support their diagnoses [5]. The consequence of unexplained misdiagnosis can be fatal in clinical practice when using an opaque classification model. Therefore, post-hoc explanations of such opaque model outcomes are important to create more confidence among clinical users of this technology.

In this paper, we investigate answers to several questions in deep learning of retinal OCT images: 1) Do we really need hefty transfer learning modules pretrained by natural images to classify OCT images?; 2) Out of several hundreds of image filters across deep convolutional layers, which ones yield the most discriminative features for OCT image classification?; 3) Can we explain the image examples misclassified by the model?; 4) Can we visualize how the deep model is learning useful features for classification? We hypothesize that OCT images can be classified using non-transfer deep learning with a better accuracy than those obtained using transfer learning.

A. Background review

AMD is one of the leading causes of blindness in adults over the age of 75 [6]. AMD can be classified as dry AMD and wet AMD. Dry AMD is clinically characterized by non-neovascular degenerative changes of the choroid, Bruch's membrane, and outer retina, the hallmark of which is drusen deposition [7]. Dry AMD can be staged based on the risk of progression to wet AMD. Choroidal neovascularization (CNV) is the hallmark of wet AMD and is caused by leakage from neovascular membranes [8]. One of the inciting causes of CNV is high levels of vascular endothelial growth factor (VEGF). Therefore, wet AMD can be treated with anti-VEGF injections which often leads to an improvement in vision. If identified late or missed, the fluid accumulation can lead to permanent damage to the photoreceptors. Therefore, early recognition is key to the management of wet AMD [9]. On the other hand, DME is characterized by fluid-filled cysts and hard exudates within the retina due to unusual leakage from damaged retinal blood vessels [10]. The diagnosis and

978-1-6654-0132-6/21/$31.00 ©2021 IEEE
DOI 10.1109/ICHI52183.2021.00020

treatment of these pathologies are made clinically based on visual acuity changes, patient symptoms, fundus appearance, and fluorescein angiography. OCT imaging is used in conjunction with these biomarkers to enhance diagnostic ability and to monitor treatment. Computer vision approaches to the classification of AMD and DME using OCT images can play an important role in aiding, automating, and optimizing clinical decisions made by an ophthalmologist. c

Recent computer vision literature has been revolutionized by the innovation of highly successful deep CNN [11]. However, deep CNN requires a large volume of labeled image samples for efficient model training that may not be available in many studies. A popular solution to the limited sample problem is transfer learning [12]. In transfer learning, an off-the-shelf deep learning model (e.g., VGG19, ResNet50) is pre-trained by millions of natural images (e.g. ImageNet data set) and then fine-tuned at the upper few layers for custom image classification task. For example, Fang et al. have used pre-trained InceptionV3 architecture with ImageNet data set for classifying macular diseases in OCT images [13]. Hwang et al. [1] have investigated the performance of three transfer learning architectures: VGG16, InceptionV3, and ResNet50 in OCT image classification, all of which have been outperformed by the deep CNN model proposed by [3]. Moreover, Mehta et al. conclude that transfer learning does not perform well on OCT images since low-level filters in transfer learning are color-dependent whereas OCT images are monochromatic [14].

We argue that pretrained models of natural images are not an intuitive choice for explaining deep learning models in medical image classification. For example, OCT images are also known as spectral domain images with black and white textures unlike time-domain natural images with three color channels. Therefore, we hypothesize that monochromatic OCT image textures are simple enough to be learned using non-transfer deep learning models to further facilitate the explainability of deep models. Kermany et al. have performed one of the earliest deep transfer learning studies in classifying OCT images with macular diseases [15]. For explainability, similar studies localize the macular regions in OCT images that are exploited by the deep learning model for classification [15], [16]. Botula et al. have proposed a lightweight CNN using only two convolutional layers for OCT Image classification [2]. They optimize their lightweight model performance by carefully visualizing and explaining the textural variations across different filter kernel sizes. Alqudah visually compares input OCT images with corresponding model responses at the last convolutional layer [3]. Das et al. have visualized their deep CNN learned OCT image features in a two-dimensional subspace to explain the effectiveness of their classifier model [17].

In line with the above studies, the contributions of this paper are as follows. First, to facilitate model explainability, we have developed a highly configurable deep CNN model without using any high-level application programmable interface (API) or pretrained transfer learning modules. Our model performance is compared with those in the state-of-the-art literature. Second, we identify a small subset of filter kernels that are sufficient for capturing macular disease patterns in OCT images in addition to providing explainability and model compression. Third, we provide clinical explanations of the misclassified images to understand the deep learning model behavior. Fourth, we visualize the features extracted from the last convolutional layer in a two-dimensional subspace for explaining the efficient feature learning capability of our classifier model.

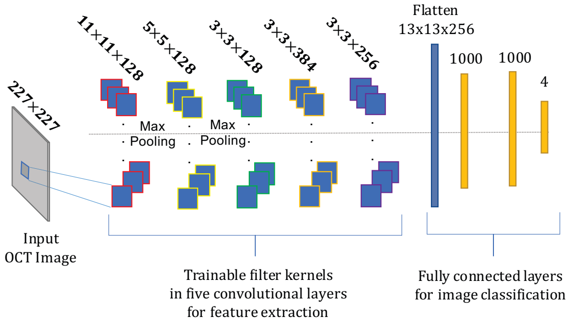

227×227
11×11×128
5×5×128
3×3×128
3×3×384
3×3×256
Max Pooling
Max Pooling
Flatten 13x13x256
1000
1000
4
Input OCT Image
Trainable filter kernels in five convolutional layers for feature extraction
Fully connected layers for image classification

Fig. 1: Proposed architecture of deep convolutional neural network for macular disease classification using OCT images.

II. Methods

We have developed, trained, validated, and tested a custom deep CNN model to conduct our proposed study. The data set, model development, and analytical procedures are discussed below.

A. OCT image data set

We use a publicly available benchmark OCT image data set sourced from [15] for macular disease classification. The data set includes 84,458 OCT image samples from follow-up studies of 4,657 patients of diverse races, ages, and gender. The image samples are clinically labeled into one of the four classes: normal (26,560 images), CNV diagnosis (37,205 images), DME (11,592 images), and drusen (8,858 images).

B. Low-level implementation of CNN:

We have used the mathematical methods and tensor data frameworks of TensorFlow v2.0 to develop a highly configurable deep CNN model from scratch to study the inner workings of such models beyond reporting of the classification accuracy. Notably, the interpretation and access to low level deep learning model parameters using popular high-level APIs such as Keras or PyTorch are often challenging and prohibitive.

C. Model architecture tuning

The proposed CNN model has been validated for a flexible architecture from three to five convolutional layers as shown in Fig. 1. The final model is selected after validating over many sets of hyperparameters. The model hyperparameters include the number of convolutional layers, the number of kernels in the five convolutional layers, the number of neurons in fully connected layers, percentage of dropout, learning rate, and regularization penalty term, and training batch size. The fully connected layers consist of two hidden layers and one output layer with four neurons for four-class image classification. All hyperparameter settings are evaluated using the validation data set after training with the training data set. Once the best hyperparameter setting is identified through validation, we train a new model with the best hyperparameter setting using both the training and validation data sets. The left-out test data set is then used to report the final test accuracy and model performance.

D. *Post-hoc explanations of deep models*

The post-hoc explanations of deep model outcomes are performed in three ways. First, we identify the CNN filters that extract the final image features to input OCT images. Since retinal OCT images involve much lesser texture complexity than natural images, a majority of the filters are expected to have no contributions to the final classification. Second, the selective filter responses at the final convolutional layer are used as features in the subsequent fully connected layers for classification. A post-hoc analysis of these selective features is performed to gain further insights into their class separability and feature distribution. We hypothesize that a deep CNN can learn the disease-specific features in retinal OCT images so well that it may not require the fully connect layers for further feature processing. Third, we identify the test OCT image samples that are misclassified by the model to further explain the reasons for misclassification with the assistance of an expert clinician. The post-hoc explainability of the model outcomes may help improve the model performance in computer-assisted diagnosis.

E. *Model evaluation*

For model validation and testing, we obtain and compare loss curves, confusion matrices, and area under the receiver operating characteristic (ROC) curves or AUCs. The classification performance of the proposed CNN model is reported on the left-out test data samples. We compare our overall classification accuracy with those reported in the literature.

III. Results

We identify a five-layer deep CNN with two fully connected layers following the validation step. The proposed model is trained entirely using OCT images without involving any pre-trained transfer learning module. All computation and analysis are performed on Google Colab, a cloud-based computing platform for deep learning and Python programming. The findings of the proposed study are summarized below.

A. *Model selection:*

The data samples are split into training (66,792) and validation (16,697) samples for tuning the model hyperparameters. The final test samples include 968 OCT images (242 images per class). The model is first trained and evaluated for 63 sets of hyperparameters using the validation data. The best model is then identified as the one with the following hyperparameter values. The numbers of filters in the five convolutional layers are 128, 128, 128, 384, 256, respectively. The number of neurons in both of the two fully connected layers is found to be 1000. The best dropout rate, batch size, learning rate, and the penalty term for the L2 regularization of the weights are identified as 0.3, 64, 0.0008, and 0.001, respectively. In the model validation with different activation functions, the scaled exponential linear unit (SELU) activation is selected in the best model. The model weights are randomly initialized using the Glorot normal distribution because of its superior performance. The best learning rate is identified by evaluating the model for an exponential function where a higher learning rate at the initial epochs gradually decays at the higher epochs. Finally, we use the best model setting to train the model using both training and validation data and report the classification performance on the test samples.

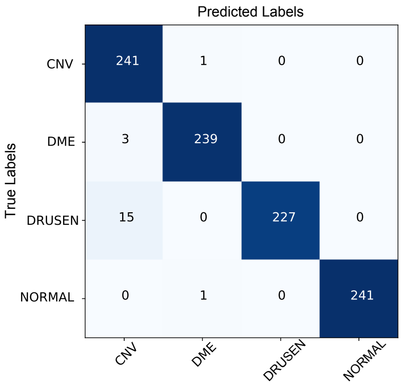


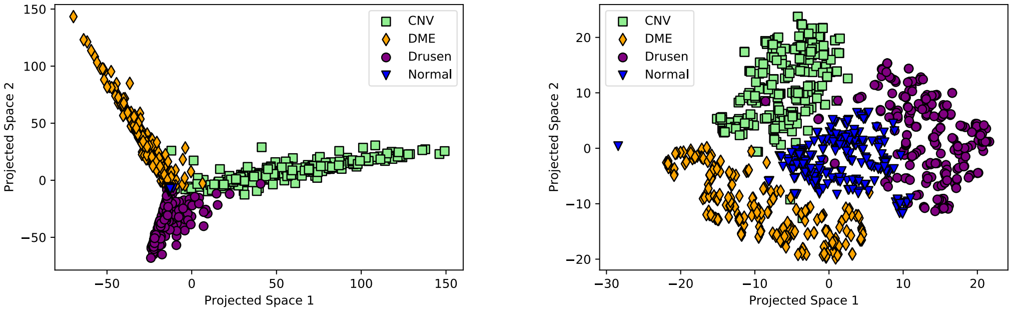
Fig. 2: Confusion matrix obtained after classifying 242 retinal OCT images corresponding to each of the four disease classes. F1-scores are 0.962 0.989 0.968 0.998 for CNV, DME, drusen, and normal, respectively. Overall sensitivity and specificity are 0.979 and 0.993, respectively.

TABLE I: Performance comparison between transfer learning and our proposed non-transfer learning approaches in classifying macular diseases using OCT images.

Reference	**Model**	**Accuracy(%)**
[13]	IFCNN	87.0
[18]	DenseNet, ResNet-50	88.0, 89.0
[16]	Localization + CNN	97.7
[2]	VGG-19	93.3
[2]	ResNet-101	96.7
[2]	LightOCT	94.5
[2]	Inception-V3	95.8
Proposed	Non-transfer CNN	**97.9**

B. *Model performance:*

The training and validation loss curves for the best hyperparameter setting suggest some overfitting. Therefore, we use an early stopping criterion to select the trained model at



(a) PCA projected

(b) t-SNE projected

Fig. 3: Visualization of CNN-derived OCT image features at the final convolutional layer in two-dimensional subspace. The normal macula samples appear at the center of the subspace. Other disease classes geometrically deviate from normal samples.

epoch 32 when the overfitting is at a minimum. The best model is therefore trained up to 32 epochs. In the model testing phase, the AUCs under ROCs for individual disease classification are found to be 0.99, 1.0, 0.99, and 1.0 for CNV, DME, drusen, and normal image classes, respectively. Figure 2 shows the confusion matrix following classification of the test image samples. The confusion matrix reveals that drusen images are mostly confused as CNV whereas the normal patient images are best classified with an 99.6% accuracy. Only 20 out of 968 test images are misclassified, which results in a 97.93% classification accuracy. In Table I, we compare our model performance with those reported in the literature which use a transfer learning approach. The comparison shows that our proposed non-transfer deep CNN model outperforms the transfer learning approaches involving pretrained ResNet, DenseNet, VGG, and Inception models.

C. *Post-hoc explanation of image features*

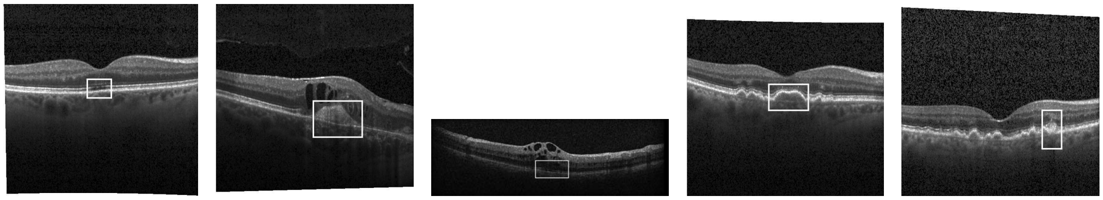
The last convolutional layer of the proposed CNN model has 256 filters. Each filter results in a 13x13 image response after convolution. We identify that only eight of the 256 filters produce non-zero responses to input images. Therefore, a single input image can be represented by a 13x13x8 (1352) dimensional feature vector. We fit a PCA model and project the data on to the two principal directions for visualization. Figure 3(a) shows that 1352-dimensional feature vectors projected using PCA shows highly discriminating patterns in only two dimensions. For visualization of the high-dimensional feature space, we use t-distributed stochastic neighbor embedding (t-SNE), which also reveals highly discriminating patterns between the four classes as shown in Fig. 3 (b). These findings reveal the relative positioning of the disease classes in the feature space with respect to normal macular images and explain why our CNN-derived image features are so effective for macular disease classification.

D. *Post-hoc explanation of misclassification*

An expert clinician visually inspected the 20 misclassified images to determine the reason behind the model's failure. Out of 20 images, the model confuses 15 drusen images as CNV. Furthermore, three DME examples are misclassified as CNV. The other two examples include normal and CNV images misclassified as DME. Figure 4 shows representative image samples for four of these misclassification pairs. The most common error is the classification between dry AMD (e.g. drusen) and wet AMD (e.g. CNV) image cases. This appears in all misclassified images except for one normal macula image, which is misclassified as DME due to the presence of an image artifact marked in Fig. 4(a). This artifact creates a skip area of the inner and outer segment (IS/OS) junction, which is possibly detected as fluid. The only misclassified image with CNV is confused as DME (Fig. 4(b)) due to the presence of an intra-retinal edema pattern similar to DME. The image samples labeled with DME are misclassified as CNV because of some shadowing on the subretinal area as marked in Fig. 4 (c). The drusen samples manifest large drusen or sometimes confluent drusen (Fig. 4(d)). Large drusen can cause elevation of the retinal pigment epithelium (RPE) layer, which may be misidentified as subretinal fluid, thereby causing the most common misclassification as CNV. A challenging drusen example is shown in Fig 4(e), which may appear as CNV to a clinician and can be accurately diagnosed using fluorescein angiography to detect leakage.

IV. Discussion

The findings of the paper are summarized as follows. First, a non-transfer deep CNN model can yield a classification accuracy similar to or even better than those obtained using transfer learning methods. For the classification of OCT images with the given sample size, transfer learning may introduce redundant depth, breadth, and complexity in the model that are not required for some medical imaging with homogeneous textures. Therefore, it is important to investigate and optimize the performance of a non-transfer deep learning model before proposing a transfer learning solution to all image classification problems. Second, low-level and non-transfer implementation of deep CNN has eased the path for explainability. Unlike transfer learning, our proposed model yields filters and filter responses that are exclusively related to retinal



(a) Normal (DME) (b) CNV (DME) (c) DME (CNV) (d) Drusen (CNV) (e) Drusen (CNV)

Fig. 4: Representative OCT images misclassified by the CNN model. The regions contributing to model confusion are highlighted by a clinician. Misclassified labels are shown in parentheses.

OCT images. A majority of the filters and filter responses appear sparse and redundant, which demands a justification of using very deep models or pre-trained transfer learning for all image classification problems. Third, we provide clinical explanations of our misclassified image samples, which can be taken into the next step of model optimization to yield near-perfect classification accuracy. Fourth, we demonstrate that selective (from only eight filters) OCT image features extracted at the last convolutional layer are highly discriminative for the four-class classification. The class separation is strongly visible even after projecting the 1352-dimensional feature vector to only two-dimension. Our class separation in appears more prominent than that found in [17] where image features are taken after the fully connected layers. This infers that our features at the final convolutional layer are quite discriminative and may not require further processing through fully connected layers prior to the classification.

V. Conclusions

In this paper, we propose a non-transfer deep learning model to demonstrate that transfer learning should not be the default choice for all medical image classification problems. Monochromatic medical images require less computational resources since we find only a fraction of all CNN filters to be useful for image classification. The visualization of high-dimensional image features reveals the effectiveness of corresponding selective image filters. Our clinical interpretation behind misclassified images explains the deep model behavior that may help in further optimizing the model performance.

Acknowledgements

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R15LM013569. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] D.-K. Hwang, C.-C. Hsu, K.-J. Chang, D. Chao, C.-H. Sun, Y.-C. Jheng, A. A. Yarmishyn, J.-C. Wu, C.-Y. Tsai, M.-L. Wang *et al.*, "Artificial intelligence-based decision-making for age-related macular degeneration," *Theranostics*, vol. 9, no. 1, p. 232, 2019.

[2] A. Butola, D. K. Prasad, A. Ahmad, V. Dubey, D. Qaiser, A. Srivastava, P. Senthilkumaran, B. S. Ahluwalia, and D. S. Mehta, "Deep learning architecture "lightoct" for diagnostic decision support using optical coherence tomography images of biological samples," *Biomedical Optics Express*, vol. 11, no. 9, pp. 5017–5031, 2020.

[3] A. M. Alqudah, "Aoct-net: a convolutional network automated classification of multiclass retinal diseases using spectral-domain optical coherence tomography images," *Medical & biological engineering & computing*, vol. 58, no. 1, pp. 41–53, 2020.

[4] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[5] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[6] N. M. Bressler, S. B. Bressler, N. G. Congdon, F. L. Ferris 3rd, D. S. Friedman, R. Klein, A. S. Lindblad, R. C. Milton, J. M. Seddon *et al.*, "Potential public health impact of age-related eye disease study results: Areds report no. 11." *Archives of Ophthalmology (Chicago, Ill.: 1960)*, vol. 121, no. 11, pp. 1621–1624, 2003.

[7] D. S. McLeod, R. Grebe, I. Bhutto, C. Merges, T. Baba, and G. A. Lutty, "Relationship between rpe and choriocapillaris in age-related macular degeneration," *Investigative ophthalmology & visual science*, vol. 50, no. 10, pp. 4982–4991, 2009.

[8] H. E. Grossniklaus and W. R. Green, "Choroidal neovascularization," *American journal of ophthalmology*, vol. 137, no. 3, pp. 496–503, 2004.

[9] J. S. Heier, J.-F. Korobelnik, D. M. Brown, U. Schmidt-Erfurth, D. V. Do, E. Midena, D. S. Boyer, H. Terasaki, P. K. Kaiser, D. M. Marcus *et al.*, "Intravitreal aflibercept for diabetic macular edema: 148-week results from the vista and vivid studies," *Ophthalmology*, vol. 123, no. 11, pp. 2376–2385, 2016.

[10] F. E. Hirai, M. D. Knudtson, B. E. Klein, and R. Klein, "Clinically significant macular edema and survival in type 1 and type 2 diabetes," *American journal of ophthalmology*, vol. 145, no. 4, pp. 700–706, 2008.

[11] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on Deep Neural Networks in Speech and Vision Systems," *Neurocomputing*, vol. 417, pp. 302–321, 2020.

[12] M. Witherow, W. Shields, M. Samad, and K. Iftekharuddin, "Learning latent expression labels of child facial expression images through data-limited domain adaptation and transfer learning," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 11511, 2020.

[13] L. Fang, Y. Jin, L. Huang, S. Guo, G. Zhao, and X. Chen, "Iterative fusion convolutional neural networks for classification of optical coherence tomography images," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 327–333, 2019.

[14] P. Mehta, A. Lee, C. Lee, M. Balazinska, and A. Rokem, "Multilabel multiclass classification of oct images augmented with age, gender and visual acuity data," *bioRxiv*, p. 316349, 2018.

[15] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[16] A. Joshi, G. Mishra, and J. Sivaswamy, "Explainable disease classification via weakly-supervised segmentation," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 2020, pp. 54–62.

[17] V. Das, S. Dandapat, and P. K. Bora, "Multi-scale deep feature fusion for automated classification of macular pathologies from oct images," *Biomedical Signal Processing and Control*, vol. 54, p. 101605, 2019.

[18] K. A. Nugroho, "A comparison of handcrafted and deep neural network feature extraction for classifying optical coherence tomography (oct) images," in *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, 2018, pp. 1–6.