Diabetic Retinopathy Classification Using a Modified Xception Architecture

Sara Hosseinzadeh Kassani
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
sara.kassani@usask.ca

Peyman Hosseinzadeh Kassani
Department of Biomedical Engineering
University of Tulane
New Orleans, USA
peymanhk@tulane.edu

Reza Khazaeinezhad
Beckman Laser Institute
University of California Irvine
Irvine, USA
Reza.k@uci.edu

Michal J. Wesolowski
Department of Medical Imaging
University of Saskatchewan
Saskatoon, Canada
mike.wesolowski@usask.ca

Kevin A. Schneider
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
kevin.schneider@usask.ca

Ralph Deters
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
deters@cs.usask.ca

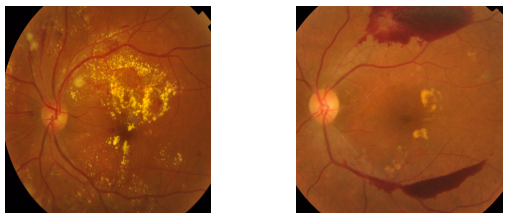
***Abstract*—Diabetic retinopathy (DR) is one of the major causes of blindness worldwide. With proper treatment, early diagnosis of DR can prevent the progression of the disease. In this paper, we present a new feature extraction method using a modified Xception architecture for the diagnosis of DR disease. The proposed method is based on deep layer aggregation that combines multilevel features from different convolutional layers of Xception architecture. The extracted features are subsequently fed into a multi-layer perceptron (MLP) to be trained for DR severity classification. The performance of the proposed approach was assessed with four deep feature extractors, including InceptionV3, MobileNet, and ResNet50 and original Xception architecture. Compared with typical Xception architecture, the aggregation of deep CNN layers can effectively fuse deep features and improve the learning process. Additionally, a transfer learning strategy and hyper-parameter tuning are adopted to further improve the overall classification performance. The performance of the proposed model was validated on the Kaggle APTOS 2019 contest dataset. Experiments demonstrate that the modified Xception deep feature extractor improves DR classification with a classification accuracy of 83.09% versus 79.59%, sensitivity of 88.24% versus 82.35% and specificity of 87.00% versus 86.32% when compared with the original Xception architecture.**

***Index Terms*—Computer-aided diagnosis, Convolutional neural network, Deep learning, Diabetic retinopathy, Transfer learning**

I. Introduction

DR is a leading cause of blindness in Type-II diabetes patients and is characterized by chronic progressive damage of the retinal microvasculature [1]. According to data provided by [2] in 2017, 9.4% of the U.S. population, about 30.3 million people, have diabetes that requires regular screening to prevent vision impairment and blindness. DR appears when blood glucose damages the retinal blood vessels. The increase of glucose level in the blood causes the arteries in the retina to weaken and leak into the eye, leading to a blur vision. At the next stage, the newly formed weak blood vessels break and leak blood into the eye, leading to vision loss. The severity of the disease is defined based on the magnitude of exudates. In the current clinical diagnosis, regular screening of diabetic patients through fundus examinations is the most effective way to detect early signs of abnormalities. If DR can be detected in early-stage and treated immediately, the occurrence of blindness can be reduced [3]. Examples of different stages of DR are shown in Fig. 1. Accurate disease grading and identification tends to be a challenging task due to the different types of fundus imaging instruments. In addition, the quality of DR fundus images may suffer from occlusion, shadow, reflection or poor illumination, making it difficult to differentiate between healthy and abnormal regions.

(a) No DR (b) Mild DR (c) Moderate DR



(d) Severe DR (e) Proliferative DR

Fig. 1: Examples of DR according to the severity of the disease from fundus images..

Over the last years, machine learning models, and in specific, deep neural networks have achieved promising results in finding hidden patterns in different computer vision tasks, especially in the field of medical image analysis [4]–[6]. Developing computer-aided diagnosis (CAD) systems based on

deep learning methods that are able to classify abnormalities can support medical decision making and improve patient care [7]. The remainder of this paper is organized as follows. Section II provides related studies in DR image classification. Section III explains the material and the proposed method. Section IV presents the experimental analysis and discussion. Finally, Section V gives the conclusions.

II. Related works

An overview of existing approaches that employed convolutional neural networks (CNNs) for DR diagnosis is presented. In a study by Zeng et al. [8], an automatic deep convolutional neural networks based on the Siamese-like architecture was adopted for DR classification. The proposed architecture accepts binocular (two fundus images corresponding to the left eye and right eye) fundus images as inputs. Zeng's approach achieved an area under ROC curve (AUC) of 95.10% and a sensitivity of 82.2%. Shanthi et al. [3] used a modified AlexNet architecture for classification of DR fundus images on MESSIDOR dataset [9] with the application of suitable pooling, softmax and rectified linear activation unit (ReLU) layers. The proposed model achieved accuracy of 96.6% on the MESSIDOR dataset. In another study conducted by Jain et al. [10], the performance of different pre-trained networks including VGG16, VGG19 and InceptionV3 were evaluated for binary and 5-class of DR classification. Different data augmentation methods were employed to balance severely skewed classes. Jain's work showed that the accuracy of models is directly related to the number of convolutional and pooling layers. The best accuracy was achieved by VGG19 at 80.40%. Hagos et al. [11] presented a DR classification model by utilizing the InceptionV3 architecture and a transfer learning strategy for small datasets. Hagos' method achieved 90.9% accuracy with an SGD optimizer and the cosine loss function for binary classification. A work by Quellec et al. [12] studied the referable lesion areas in DR images using a weakly-supervised model and a generalization of the backpropagation method to generate heatmaps. Quellec's proposed approach achieved an area under the ROC curve (AUC) of 95.50% on Kaggle dataset and 94.90% on E-Ophtha dataset. Gargeya et al. [13] employed an automated feature-learning algorithm for the assessment of glaucoma. The performance of the model was validated on the two public, MESSIDOR 2 and E-Ophtha databases. Gargeya's proposed algorithm achieved a 97.00% AUC with 5-fold cross-validation. Finally, in a study by Orlando et al. [14], an ensemble of deep learning models was employed for red lesions detection in fundus images. In that method, first, patches of 32×32 pixels were extracted and fed into a Deep CNN. Additionally, hand-crafted features were extracted and passed to a random forest (RF) classifier. Orlando's method highlighted that the hybrid feature vector of both deep learning-based and hand-crafted features could improve the performance of the networks and achieve 89.32% AUC.

With the above review in place, the motivation of this paper is to build an automatic diagnosis system for DR severity classification from fundus images with the advantage of inception module, residual blocks and separable convolution layers. Considering the large variation of lesion structures, employing hand-crafted features, which are often based on low-level features, are computationally intensive and requires elaborate fine-tuning that could introduce complexity to the model. Extensive task-specific pre-processing and data augmentation techniques make the model susceptible to noise and artifacts for large scale image analysis. Employing very deep architectures for small data samples also could have the issues of poor local minima and gradient vanishing. To address the above limitations, we created a novel framework based on the deep layer aggregation [15] method which can effectively classify the severity of the disease to their respective classes. The aggregation method in the proposed architecture can fuse possible deep features related to DR structures.

The contributions of this paper are summarized as follows:

- First, we demonstrate that a proper set of features extracted from a relatively small dataset without data augmentation techniques can achieve promising performance. In this work, with a view to addressing the issue of a low detection rate on a small dataset; we applied both L1 and L2 regularization to prevent overfitting and improve the generalizability of the proposed modified Xception deep extractor.
- Second, we employ a modified Xception architecture to boost the discriminative capability of high-level information extraction that could improve the recognition of slight differences between classes in DR severity classification. We prove that the combination of Xception network components and the aggregation of intermediate CNN layers that acts as auxiliary supervision can lead to a better detection rate despite the limited number of training samples. We use an MLP classifier for training the extracted features. We also compare our modified Xception deep extractor to the original Xception, InceptionV3, ResNet50 and MobileNet architectures to demonstrate the benefits of the proposed approach.
- Finally, transfer learning using pre-trained deep CNN and hyper-parameter tuning are key components in the training process and proved to be very useful for medical image analysis. Instead of training a model with randomly initialized weights, we use the weights from weights trained on ImageNet dataset as weights initialization. We demonstrate that our proposed classification model obtains better results compared with deep CNN models. The performance of the proposed model is evaluated in terms of accuracy, sensitivity and specificity.

III. Materials and Methods

A. Proposed Approach

In this study, we design a classification framework by a modified Xception architecture. We choose Xception as the backbone network for deep feature extraction. This architecture is composed of three different components; the inception

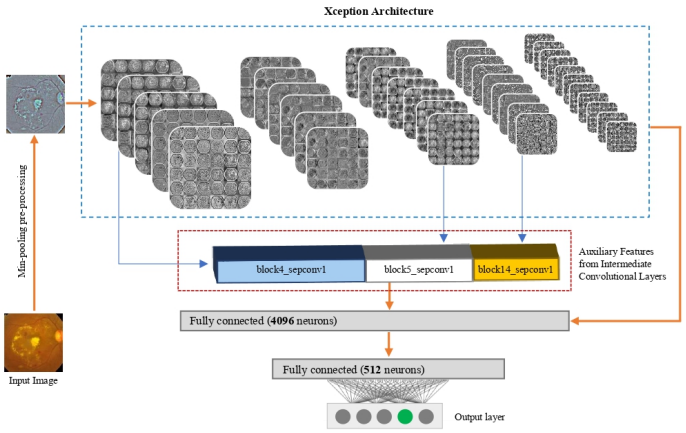


Fig. 2: Illustration of the proposed network architecture

module, depthwise separable convolution layers and residual blocks. In addition to these components, in our framework, we also improve the performance of DR severity classification by stacking high-level features extracted from intermediate convolutional layers. To achieve this, first, we omit the top layer of the Xception network to transfer pre-trained knowledge from ImageNet dataset and then concatenate features extracted from intermediate CNN layers. This is the main difference between our method and the original Xception architecture. Also, in the Xception architecture, the standard convolutional layers are replaced with depthwise separable convolution layers that help to reduce the computational complexity, which is useful when implementing real-time applications. In addition to the modified Xception deep extractor for this task, we employed the min-pooling [16] image pre-processing technique to aid in the features extraction and further improve the classification performance. The illustration of our proposed deep learning network is shown in Fig 2.

The performance of the proposed method is compared with the InceptionV3, MobileNet, and ResNet50 and original Xception architectures. The network structure of the selected architectures are as follows:

- **Xception:** The Xception architecture [17], introduced by Francois Chollet, is an extension of the Inception architecture. This architecture is a linear stack of depthwise separable convolution layers with residual connections. The depthwise separable convolution aims to reduce computational cost and memory requirements. Xception has 36 convolutional layers structured into 14 modules, all include linear residual connections, except for the first and last modules. The separable convolution in Xception separates the learning of channel-wise and space-wise features. Also, the residual connection introduced by He et al. [18] helps to solve the issues of vanishing gradients and representational bottlenecks by creating a shortcut in the sequential network. This shortcut connection is making the output of an earlier layer available as input to the later layer using a summation operation rather than being concatenated.
- **MobileNet:** MobileNet [19], developed by the Google research team, consists of depthwise separable convolution layers. A high accuracy rate can be achieved in MobileNet architecture with a small number of hyperparameters. Depthwise separable convolution layers map the cross-channel correlations and spatial correlations in the feature maps of input images. A depthwise separable convolution combines two major components: a depthwise convolution and a pointwise convolution. Depthwise convolution applies a single spatial filter for each input feature map, and then pointwise convolution (1×1) uses a filter for cross-channel patterns. Standard convolutional layers capture the spatial patterns and cross-channel patterns simultaneously while a separable convolutional layers deal with spatial patterns and cross-channel patterns separately.
- **InceptionV3:** The Inception module proposed by Szegedy et al. [20] consists of 42 layers. The InceptionV3 is the third generation of Inception module proposed by the Google Brain and consists of 159 layers in total. The main idea of the Inception module is to modify small kernels with the large kernel to learn multi-scale representations and reduce the computational complexity and the total number of parameters.
- **ResNet50:** The concept of a residual block is introduced in deep residual learning network (ResNet) [18]. Residual blocks are designed to add a connection from the input of the first block to the output of the second block. This adding operation helps the residual block learns the residual function and avoids parameter explosion. ResNet50 architecture is a 50 layer residual blocks consisting of a convolutional layer, 48 residual blocks, and a classifier layer with small filters size of 1×1 and 3×3. This architecture won the ILSVRC 2015 classification task and achieved very promising results on ImageNet and MS-COCO object detection competitions.

B. Dataset description

The dataset used for this research is the APTOS 2019 on diabetic retinopathy classification contest organized by the Asia Pacific Tele-Ophthalmology Society available at [21]. The goal of this challenge is to develop machine learning models to automatically screen fundus images for early detection of DR in rural areas where conducting medical screening is time-consuming and difficult. The dataset consists of a total of 3662 retina images collected from multiple clinics under a variety of imaging conditions using fundus photography from Aravind Eye Hospital in India. Fundus images provided in this dataset are categorized into five classes: No DR (Class 0), Mild DR (Class 1), Moderate DR (Class 2), Severe DR (Class 3), Proliferative DR (Class 4). As demonstrated in Table I, the class distribution of APTOS dataset is highly imbalanced, i.e. 49%, 8% and 5% for the No DR, Proliferative DR and Severe DR classes, respectively.

TABLE I: The class distribution of APTOS dataset.

Label	Count
No DR	1805
Mild DR	370
Moderate DR	999
Severe DR	193
Proliferative DR	295

C. Data pre-processing

The fundus images of the provided dataset were collected from various clinics with different cameras. The acquired input images have considerable variation in image intensity. Thus, we performed several pre-processing methods to optimize the training process.

- **Resizing:** Concerning the variations of the dimension of the original images, we resized all images to 819×614 according to the aspect ratio and then cropped them from center to the final resolution of 600×600 pixels using bicubic interpolation to ensure each retinal circle is located at the center of the image.
- **Min-pooling pre-processing:** With a view to enhance the clarity of blood vessels and lesion areas, we employed the method introduced by Graham [16]. According to this method, the black pixels from the background are first removed, and then the image normalization method based on the min-pooling filtering is conducted as given by:

$$I_c = \alpha I + \beta G(\rho) * I + \gamma \tag{1}$$

  where * denotes the convolution operation, I denotes input image, and $G(\rho)$ denotes the Gaussian filter with a standard deviation of ρ. And, α, β, γ are pre-defined parameters. Fig. 3 shows examples of the original image and the corresponding pre-processed image using min-pooling filtering method.
- **Image normalization:** The intensity values of cross-channel of all images were normalized from [0, 255] to [−1, 1]. This operation helps to remove bias from the features and achieve a uniform distribution across the dataset. We also normalized images using ImageNet mean subtraction as a pre-processing step. The ImageNet mean is a pre-computed constant derived from ImageNet database [22].

D. Metrics for performance evaluation

The performance of the all classifiers is assessed based on three evaluation metrics, namely, accuracy, sensitivity and specificity. Given the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), these measures are mathematically expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{2}$$

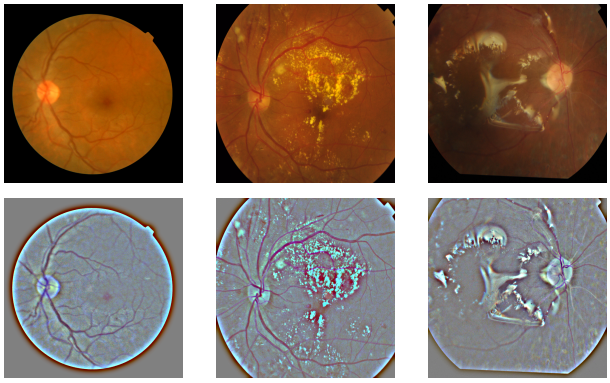


Fig. 3: Examples of original fundus images (top row), and corresponding processed images (bottom row).

$$Sensitivity = \frac{TP}{TP + FN} \times 100 \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \times 100 \tag{4}$$

IV. Experiment and Results

A. Experimental Setup

Our policy for data separation is as follows, 70% of the images of each class are assigned to the training set, 20% to the validation set, and the remaining 10% to the test set. For min-pooling pre-processing, α, β, ρ and γ were set to 4, -4, 10, 128, respectively. Learning rate, $\beta 1$ and $\beta 2$ were set to 0.00001, 0.3 and 0.6 for Adam optimizer, respectively. The MLP classifier with two fully connected networks of 4096 and 512 hidden neurons was employed for training. We utilized ReLU activation function for fully connected layers and L1 and L2 regularization with values of 0.00001 and 0.000001, respectively. Dropout in the last fully-connected layer is set to a rate of 0.5 to prevent over-fitting. The batch size is set to 128, and all models are trained for 1000 epochs. Our experiment is implemented in Python using the Keras package with Tensorflow as backend and is run on Nvidia GeForce GTX 1080 Ti GPU with 11GB RAM.

B. Results and discussions

All models are trained on 2657 images, validated on 662 images, and results are obtained from 343 test images which were not used in the training and validation phase. The accuracy, sensitivity and specificity of the obtained results are summarized in Table II. The experimental results in Table II confirm that the proposed model with multi-layer feature fusion, by aggregating features from intermediate convolutional layers outperforms all counterparts and achieves the highest accuracy for three classification metrics.

Referring to Table II, we observe that the proposed model delivered high accuracy (83.09%) without artificially augmented dataset. High sensitivity (88.24%) and high specificity (87.00%) result are also obtained by the proposed model. To

TABLE II: Classification results from pre-trained networks and proposed architecture. The bold value indicates the best result; underlined value represents the second-best result of the respective category

Method	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Xception	<u>79.59</u>	<u>82.35</u>	<u>86.32</u>
InceptionV3	78.72	63.64	85.37
MobileNet	79.01	76.47	84.62
ResNet50	74.64	56.52	85.71
Proposed model	**83.09**	**88.24**	**87.00**

justify the performance of the proposed architecture, the performance of ResNet50, InceptionV3, Xception and MobileNet architectures are evaluated and compared with the proposed model. As shown in Table II, our model improves ResNet50 up to 8.45%, InceptionV3 up to 4.37%, MobileNet up to 4.08% and the original Xception architecture up to 3.50% in terms of accuracy, which is considered significant.

Moreover, the original Xception architecture (79.59%) gives a better performance than ResNet50, InceptionV3 and MobileNet architectures. This is probably because of the benefit of the combination of depth-wise and point-wise blocks, inception modules and residual blocks in Xception architecture compared to depth-wise and point-wise convolutional layers in MobileNet, inception modules in InceptionV3 and residual blocks in ResNet50 architecture. The worst classifier is ResNet50 with an accuracy of 74.64%, sensitivity of 56.52%, and specificity of 85.71%. The obtained results indicates that the role of combination of these modules together is useful. Additionally, including auxiliary features by aggregating intermediate convolutional layers from Xception architecture can further help to extract more discriminative features.

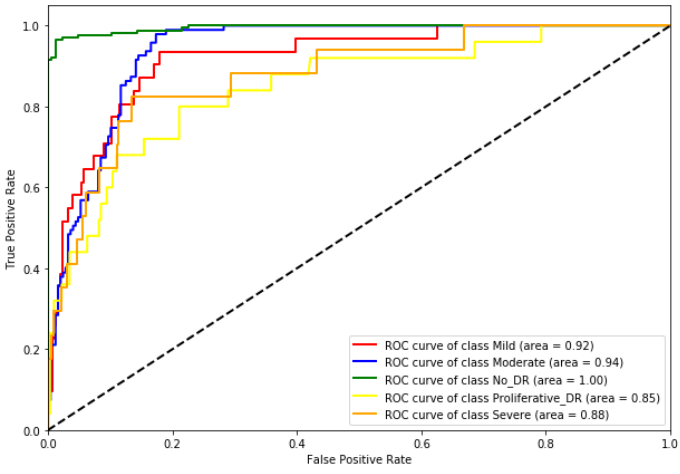


Fig. 4: . ROC curve of the proposed model on APTOS dataset for DR screening. The true positive rate represents sensitivity, and false-positive rate represents 1– specificity.

The ROC curve of the proposed model is shown in Fig. 4. Referring to this figure, we obtain AUC of 100%, 94.00% and 92.00% for No DR, Moderate, and Mild classes, respectively. However, the Severe DR and Proliferative DR gain 88.00% and 85.00%, respectively. Intuitively, this low detection rate could be explained by the imbalance training sample with 5% and 8% for these two classes. Also, the morphology of some of the fundus images has anatomic heterogeneity and shape variation that may significantly affect the identification of pathologic structures.

V. Conclusion

In this study, we developed a novel CNN model based on a modified version of Xception architecture and with aggregation of deep CNN layers for DR severity classification. The proposed method can effectively fuse feature maps of different depths and provide an accurate and computationally efficient approach for DR severity classification. Min-pooling pre-processing is employed to improve the color contrast of input images. The proposed model is trained without data augmentation. In addition, to tackle the highly imbalanced classes in the dataset, two strategies, i.e., L1 and L2 regularization was employed which help to overcome overfitting issue and hence improve the model generalizability to the new unseen data. The modified Xception deep extractor achieved a much better performance on the APTOS dataset compared to the original Xception architecture and a few other state-of-the-art algorithms. We experimentally proved that the combination of Xception network components and the aggregation of intermediate CNN layers lead to promising performance despite the limited number of training images. For future research direction, we investigate the applicability of other pre-trained CNN models. Moreover, the employed training data samples are limited. Hence, extending the dataset size by additional data sources, may also lead to better results.

References

[1] T. J. Jebaseeli, C. A. D. Durai, and J. D. Peter, "Segmentation of retinal blood vessels from ophthalmologic diabetic retinopathy images," *Computers & Electrical Engineering*, vol. 73, pp. 245–258, 2019.

[2] "National Diabetes Statistics Report." [Online]. Available: https://www.cdc.gov/diabetes/data/statistics-report/index.html

[3] T. Shanthi and R. Sabeenian, "Modified alexnet architecture for classification of diabetic retinopathy images," *Computers & Electrical Engineering*, vol. 76, pp. 56–64, 2019.

[4] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Breast cancer diagnosis with transfer learning and global pooling," *arXiv preprint arXiv:1909.11839*, 2019.

[5] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, R. Deters *et al.*, "A hybrid deep learning architecture for leukemic b-lymphoblast classification," *arXiv preprint arXiv:1909.11866*, 2019.

[6] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Classification of histopathological biopsy images using ensemble of deep learning networks," *arXiv preprint arXiv:1909.11870*, 2019.

[7] S. H. Kassani and P. H. Kassani, "A comparative study of deep learning architectures on melanoma detection," *Tissue and Cell*, vol. 58, pp. 76–83, 2019.

[8] X. Zeng, H. Chen, Y. Luo, and W. Ye, "Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30 744–30 753, 2019.

[9] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed database: the messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014. [Online]. Available: http://www.ias-iss.org/ojs/IAS/article/view/1155

[10] A. Jain, A. Jalui, J. Jasani, Y. Lahoti, and R. Karani, "Deep learning for detection and severity classification of diabetic retinopathy," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*. IEEE, 2019, pp. 1–6.

[11] M. T. Hagos and S. Kant, "Transfer learning based detection of diabetic retinopathy from small dataset," *arXiv preprint arXiv:1905.07203*, 2019.

[12] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Medical image analysis*, vol. 39, pp. 178–193, 2017.

[13] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[14] J. I. Orlando, E. Prokofyeva, M. del Fresno, and M. B. Blaschko, "An ensemble deep learning based approach for red lesion detection in fundus images," *Computer methods and programs in biomedicine*, vol. 153, pp. 115–127, 2018.

[15] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2403–2412.

[16] B. Graham, "Kaggle diabetic retinopathy detection competition report. 2015," in *URL https://kaggle2. blob. core. windows. net/forum-message-attachments/88655/2795/competitionreport. pdf*.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[21] "APTOS 2019 Blindness Detection." [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection/

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.