Ocular Diseases Recognition from Fundus Images with Convolutional Neural Networks

Guillaume Thiry, Jinyan Tao, Alexandre Bense

February 7, 2021

Link to the code

Abstract

Addressing medical problems with Deep Learning methods has become a common practice in Computer Vision in recent years. There are flourishing discoveries and new algorithms that help solve problematics in modern medicine. In our work, we focused on a disease classification problem using fundus images, emphasizing implementation of modern techniques in data augmentation as well as eligible interpretations of decisions made by our algorithms. In the end, we were able to demonstrate the efficiency of these state-of-the-art methods on the disease recognition task and their potential to answer crucial questions in medicine.

1 Introduction

According to the World Health Organization (WHO), more than 1 billion people in the world have a vision impairment that could've been prevented, induced by diseases such as cataract, diabetes or glaucoma. The early detection and treatment of these diseases have therefore become a health priority. In the meantime, the progress in automatic visual recognition led by cutting-edge Deep Learning algorithms encourages researchers to develop models to automatically detect these diseases only using pictures of the eyes.

ODIR-5k is a labeled dataset created in 2019 containing color fundus of the left and right eyes of 5000 patients with cataract, diabetes, glaucoma, myopia and so on. In total, 8 different labels (sometimes combined) describe the images in the dataset. Some research has already been conducted on this dataset, with solutions of different complexities and different successes (see comparison in Section 4.2).

Several attempts have been made to approach the ocular disease recognition with deep neural networks and some of them achieved high accuracy in predicting diabetes, glaucoma and hypertensive retinopathy [1, 2, 3]. However, these works were mostly designed for a binary classification task, with only 2 labels. Moreover, no detailed interpretation for the model decision was reported to our knowledge. In our project, we implemented Deep Learning techniques such as Convolutional Neural Networks (CNNs) and complex pretrained ImageNet architectures to solve a multi-class classification problem on this eight-label dataset. In addition, we interpreted the CNN decisions by localizing discriminative regions on images with Gradient-Weighted Class Activation Mapping (Grad-CAM), regions of interest disclosed by Grad-CAM aligned with rationales behind ophthalmological diagnosis.

2 Models and Methods

2.1 Data augmentation

Data augmentation is a well-known set of techniques used to improve the generalization capacity of the model by giving it more images during the training. These new images are usually generated by modifying the already available images to a certain extent.

2.1.1 Traditional methods

The mainstream approach for data augmentation is to apply simple functions to each original image to create new ones. To do so, we used three simple traditional data augmentation methods as described in [4]. The first one applied a random Gaussian noise of variance 0.01 (which improves the robustness of the network [5]), the second imposed a random rotation between $-\frac{\pi}{4}$ and $\frac{\pi}{4}$ and the third one rescaled the intensity between its 0.2 and 99.8 percentile to enhance the contrast. Different from Md Islam et al., [6], flipping pictures of eyes wasn't applied due to the asymmetric structure of the left and right eyes.

2.1.2 Deep Convolutional Generative Adversarial Networks

Moreover, we performed a more sophisticated method which relies on deep convolutional generative adversarial networks (DCGAN) which has already showed promising applications for medical images classification [7] and is famous for its enhanced robustness. This method has

already been used for retinal images [8, 9] to generate fake images with glaucoma.
For the implementation we used the traditional DCGAN architecture [10] which has been employed to generate fake retinal images. We only changed the activation function from the Generator from ReLu to LeakyReLu to solve convergence issue, presumably due to a vanishing gradient. It is essential to notice that in this architecture, the images are resized to 64*64 pixels before the training. To know which label we had to apply to the generated images, we had to split the data and train a DCGAN for each class respectively. This structure was implemented using PyTorch.

2.2 Deep learning algorithms

The Convolutional Neural Networks (CNN) are being widely-used and encouraged in the field of Computer Vision. In our work, we tested several implementations of this model using PyTorch, from the simple CNN to more evolved architectures such as VGG and Inception [11, 12].

2.2.1 Simple Convolutional Neural Network

We first implemented a standard form of CNN. The first part of the network is composed of different layers linked with convolutions and max-pooling functions. To prevent overfitting, batch normalization and dropout functions are used between the layers. In this part, the layers' width decrease while their depths increase. Then, the second part of the network consists of dense layers, smaller and smaller, linking the last convolutional layer to the final output layer used for the prediction. All the layers in this network are activated by the ReLU function.

The training was done by small batches of images, over many epochs. The training data was undersampled as some classes were too prominent, causing problem of generalization for the least frequent classes. To assess each prediction, we used the cross entropy loss for multi-class classification, with a weight decay parameter adding a $\mathcal{L}^2$-regularization. Finally, the optimization of the network's parameters was performed with Stochastic Gradient Descent (SGD) in respect to this loss.

After having tried several versions of this network's architecture, the final arrangement is the one presented in Table 1.

2.2.2 Pretrained ImageNet Models

In addtion to the standard CNN, Very Deep Covolutional Networks (VGG13 with batch normalization)[11] and Inception v3 by Google [13, 12] were implemented. Pretrained on the ImageNet data, these models were finedtuned with all possible parameters on our training data. Similar to the previous CNN model, a binary cross entropy loss was applied for the multi-class classification problem, and was optimized with SGD.

Function	Size
Input	$512 \times 512 \times 3$
Convolution_1	$500 \times 500 \times 12$
Batch_Norm_1	$500 \times 500 \times 12$
Convolution_2	$500 \times 500 \times 36$
Batch_Norm_2	$500 \times 500 \times 36$
Max_pooling_1	$125 \times 125 \times 36$
Dropout_1	$125 \times 125 \times 36$
Convolution_3	$120 \times 120 \times 72$
Batch_Norm_3	$120 \times 120 \times 72$
Convolution_4	$120 \times 120 \times 144$
Batch_Norm_4	$120 \times 120 \times 144$
Max_pooling_2	$30 \times 30 \times 144$
Dropout_2	$30 \times 30 \times 144$
Dense_1	4096
Dense_2	512
Dense_3	64
Dense_4	8

Table 1: Layers of the proposed CNN

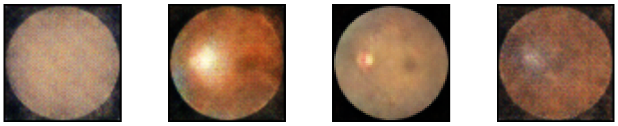
2.3 Model Intepretability

To interpret the decisions made by the CNN for the classification task, Gradient-weighted Class Activation Mapping (Grad-CAM)[14] was implemented to illustrate discriminative localization on images for each class. The Grad-CAM scheme follows the workflow where it firstly forward-propagates an image of a given label through the network to obtain an output y^c . It then back-propagates to a certain convolutional layer to obtain the gradients with respect to feature map activations and assign neuron importance weights. Lastly, it returned the 2D localization coarse heatmap by a weighted combination of forward activation maps.

3 Results

3.1 Data Augmentation

After a successful implementation of traditional data augmentation, the DCGAN was employed. However, due to the dataset imbalance, it was only able to converge on the two mostly represented classes: the normal (N) and the diabetes (D). We can see the generated images at different epochs on Figure 1. For these classes, the results are promising as the generated images highly resembles the real retinal images. After a sufficient training of nearly 100 epochs, we observed the appearance of the optical disc and the macula on the image. We noticed at epoch 200, the vein-like shape merges around the optical disc area. Nevertheless, as we resized the images to 64*64 pixels (compared to 512*512 originally), the resolution is reduced by a factor of 8 and the generated images thus are not perfectly accurate. For example, the optical disc



D:Epoch 30 D:Epoch 100 D:Epoch 200 N:Epoch 100

Figure 1: Example of generated image using DCGAN for diabetes (D) and normal (N) eyes

remains quite blurry even at epoch 200 and therefore it seems very unlikely that these images will provide additional information for the detection of glaucoma [8, 9]. The same conclusion can be drawn for hemorrhages that cannot be represented on low-resolution images but are essential for the detection of diabetes [15].

3.2 Learning with Data Augmentation

We compared the different strategies of data augmentation on the binary classification 'Normal vs. Diabetes'. To do that, we trained the same CNN during 50 epochs on the original data (*default*), the original data + the classical augmented data (*augmented*) and the original data + the GAN augmented data (*gan*). A small (15 percent) portion of the data was held out as the test dataset. Two metrics were used after each epoch to assess the classification power. The first one, the F1-score, is a well-used evaluation metric for imbalanced dataset (here, we have more 'normal' patients than 'diabetes' ones). The second one, the recall, discloses the true positive rates for diabetes prediction. Both scores for the three scenarios are shown in Figures 2 & 3.

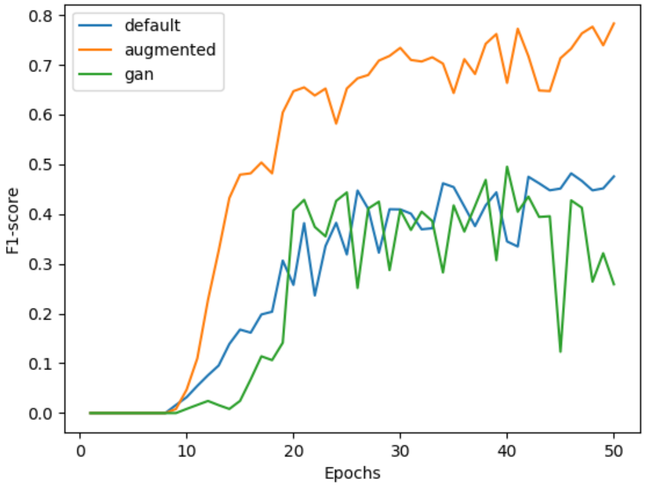


Figure 2: F1-score for different data augmentation

As expected, having more data thanks to data augmentation gave better results and faster convergence since the model have more data to train and generalize. However, it is not the case with the GAN-produced images, with which the model learned slower than the default one, and eventually gave similar accuracy but not better. This might result from the quality of the images generated by GAN, a problem that is discussed in Section 4.1.

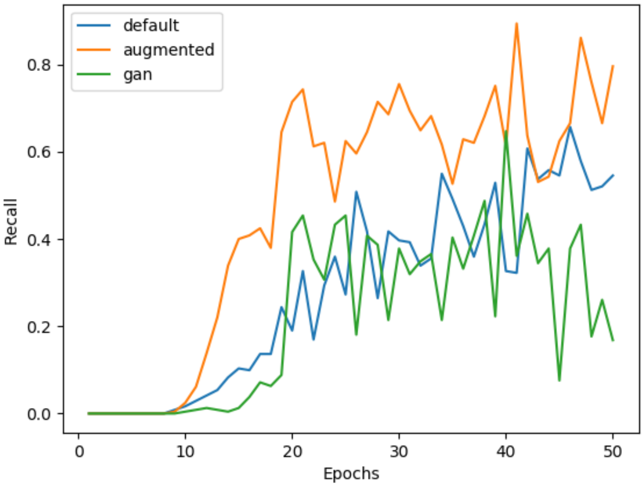


Figure 3: Recall for different data augmentation

3.3 Learning with Pretrained Computer Vision Models

In order to improve the accuracy of our classification task, we experimented two other robust image recognition models VGG [11] and Inception v3 by Google [12, 13]. Based on the data augmented by classical methods, the prediction accuracy on the test dataset was recorded for 50 epochs (see supplementary for details) and compared with our CNN model. The accuracy of the best-performed models in 50 epochs were benchmarked by recall per label (see Figure 4). Among the three models, VGG achieved the highest overall F1-score, slightly outperformed Inception v3 while CNN achieved F1-score of 0.45 (Table 2). Pretrained on ImageNet and fine-tuned via training on our dataset, this transfer learning approach slightly increased the accuracy compared to the classical CNN. Further improvement of the accuracy is still conceivable, especially by application of ensemble methods. However, it is worth noting that the classical CNN did better on almost every disease while the pretrained models were especially good at recognizing the 'normal' class. Furthermore, in the cases of disease diagnosis, false positives are more tolerable than false negatives with the purpose of preventing disease progression and late treatment, therefore, our CNN shows advantages.

Table 2: Model performance by F1-score

Image recognition model	F1-score (micro)
CNN	0.45
VGG13 (with batch normalization)	0.55
Inception v3	0.52

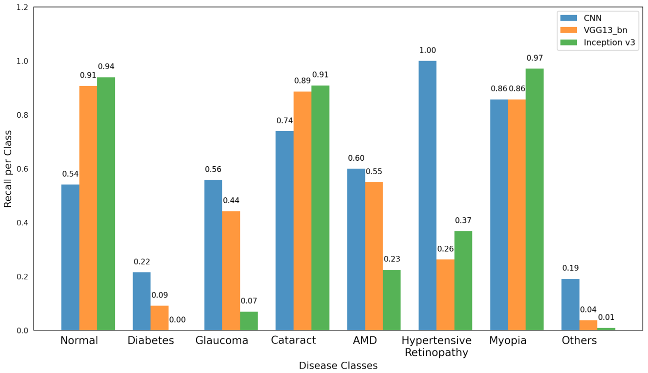


Figure 4: Recall score per class of best-performed models during training with CNNs, VGG and Inception v3 respectively

3.4 Learning with Meta Data

In addition to image inputs, meta data including patient age and gender can be considered as additional features for our models. A quick logistic regression using these two features gives a score of 0.44 to predict the labels, showing that even if there is not a very strong correlation, adding those could help in some way. However, their addition to a CNN model can be tricky: we tried to add the normalized age and the digit-encoded gender to our CNN by concatenation on the 3rd dense layer to flattened image features. It was then passed through two additional dense layers and gave a final F1-score dropping to 0.12. This shows that a more careful model architecture design should be considered to integrate the meta data, without however any certainty of success.

3.5 Interpreting Decision of CNN with Grad-CAM

To interpret the decisions of the CNN, we created 'visual explanations' of the discriminative regions on images for each label with Grad-CAM [14]. In essence, we combined the original picture with the gradient-weighted heatmap to indicate the localization of CNN's interest for disease recognition in Figure 5.

In general, Grad-CAM results showed that discriminating regions are around the optic disc, blood vessels, pigmentation and retinal deposits. For Glaucoma, the decisive regions given by Grad-CAM lies around the area of the optic disc, corresponds to patho-physiological enlargement of the optic disc in the optic assessment [1]. With Cataract, a class with high predictability, Grad-CAM showed most of the pixel contributes equally with respect to the gradients. For their part, AMD and HR are largely characterized by yellow deposits under retina, which are accurately localized by the heatmap [16, 2]. For diabetic retinopahy, the fundus shows hemorrhage that are detectable (see Figure 5c); however for minor retina blood vessel distortions present in most cases of diabetes, it is tricky to detect, which led to misclassifications with normal fundus [3]. The last class, named 'Other', is composed of many different less frequent diseases, making it really challenging to have an accurate classification and a meaningful heatmap. Finally, the easiness to localize discriminating regions reflects well on the classification scores of our models: for the classes that are the easiest to detect, the decisive regions for disease recognition are clear and align with criteria in ophthalmologic assessments. On the other hand, reasons behind a classification between a normal and diabetes fundus remained ambiguous, and the final scores are lower.

It is also important to note that Grad-CAM has its disadvantage: it only shows gradient-weighted heatmap on height and weight dimensions, but forbids interpreting decisive effects of color channels for disease recognition. For example, myopia is characterized by redness and cataract is characterized by white fundus images.

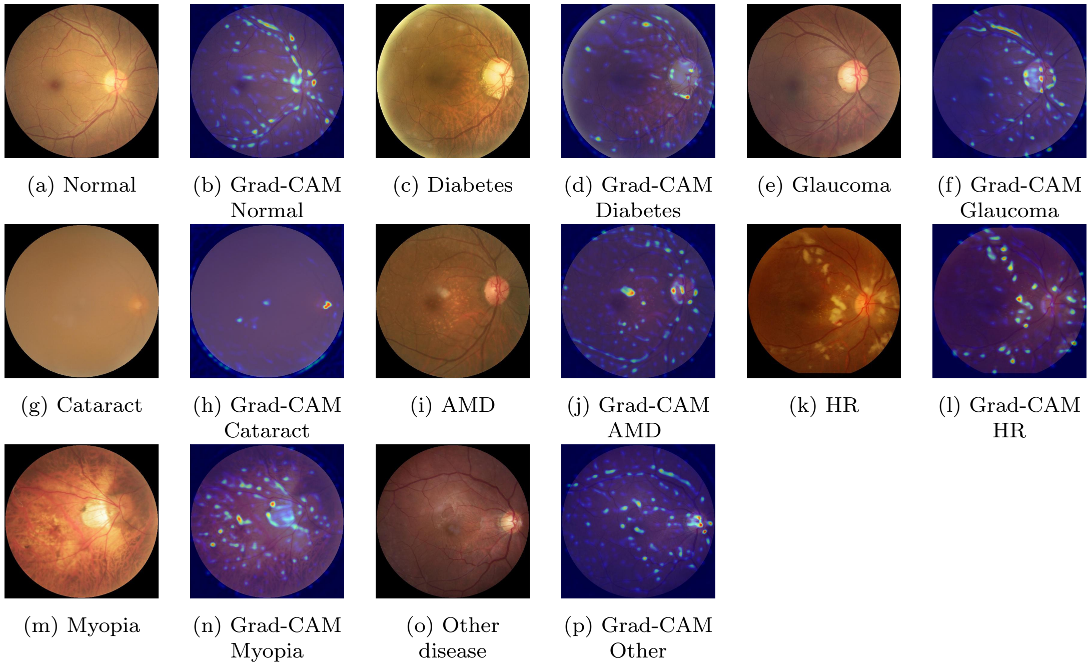
4 Discussion

4.1 GAN-based preprocessing

The GAN approach for data augmentation still remains interesting and promising even though we did not observe a positive add-on to our prediction task. Indeed, the DCGAN only converged for the two mostly represented classes, the normal (N) and the diabetes (D), and we did not find a convenient and simple solution to this problem. To deal with that, a solution using the CycleGAN method instead could generate fake images for all classes using the original normal and diabetes images. Yet, we did not succeed to implement this method using the code available on GitHub repositories due to their reproducibility issues (https://github.com/aitorzip/PyTorch-CycleGAN; https://github.com/yunjey/mnist-svhn-transfer).
In any case, the images generated by the DCGAN method as we can see on Figure 1 are already encouraging even if they are lacking precision to correctly describe the details needed for the classification. To solve this issue, we need to keep the full resolution of 512*512 pixels of the images without resizing them to 64*64, which requires changing the architecture of the generator and the discriminator networks, adding 3 additional layers of convolution, batch normalization and activation. On the other hand, this solution will add a lot of weights to optimize and will therefore slow down the training.

4.2 Improving Model Performance

The results given by our different models are promising, even if we have not been able to outperform all the known baselines for this dataset. First if we compare with [17], an simple article using a standard CNN structure to explore the dataset, we see that we have been able to match the results using the same type of network (with a F1-



(a) Normal (b) Grad-CAM Normal (c) Diabetes (d) Grad-CAM Diabetes (e) Glaucoma (f) Grad-CAM Glaucoma

(g) Cataract (h) Grad-CAM Cataract (i) AMD (j) Grad-CAM AMD (k) HR (l) Grad-CAM HR

(m) Myopia (n) Grad-CAM Myopia (o) Other disease (p) Grad-CAM Other

Figure 5: Original images of 8 classes and their Grad-CAM localizations. Grad-CAM localization was specifically visualized on the 4th convolutional layer in CNN. AMD refers to age-related macular degeneration and HR to hypertensive retinopathy. Regions of interest are colored by importance, from green to red.

score around 0.45) and even to beat this score with more advanced networks (VGG and Inception). However, if we compare our results to [6] or [18], we can see that a F1-score around 0.8 could be expected with a more robust model.

Nevertheless, this was not the core of our work as we focused on other aspects such as the interpretability of the network. Anyhow, to maximize our score in a future work, some other techniques and methods could be explored. First, the more training data the better. Therefore, implementing even more methods of data augmentation (potentially retrying one using GANs) could be a point in case data collection is not possible. Another possibility could be to try other networks architectures and to spend an extended time to fine-tune every possible parameters. Finally, a common practice in data challenges such as [18] is to use ensemble methods, averaging the results on many different models. This method is very time-consuming but also very effective to increase the score of a model and prevents overfitting issues [19].

4.3 Improving Model Interpretability

For the model interpretation, due to the subtle differences among ocular diseases on images (blood vessel distortion, hemorrhage, optic disc abnormalities etc.) and the lack of an obvious objects to be recognized, image segmentation was excluded from applicable interpretation methods. Though Grad-CAM helped discover critical class-specific features, there are other approaches to be investigated to improve the feature explanation. In our presentation, the coarse heatmap was used, however, there exists guided Grad-CAM which allows a better resolution by fusing Guided Backpropagation with Grad-CAM [20]. Similarly, SmoothGrad can be applied to identify influential pixels as well [21], which might increase resolutions on the final presentation of discriminative regions.

5 Summary

To conclude, we explored the dataset in as many ways as possible in the intriguing context of machine learning applications in medical images. While trying our best to build the best predictor using Convolutional Neural Networks, we also focused on more innovative approaches, one trying to perform data augmentation using GANs, and the other trying to give a meaningful interpretation of the trained network.

Having showed promising preliminary results, the data augmentation by GANs could, with more data and flexibility in time, become a truly reliable tool to increase the amount of training data for images, as it's already been highly implemented in simpler tasks. Moreover, the results given by the model-interpreting method convinced the untrained eyes and should be furthermore challenged by medical professionals, to testify the reliability of the underlying understanding from our model on these images. If the reliability is confirmed, applications of deep learning models in solving similar tasks could largely depict a promising prospect in AI-assisted disease diagnosis and progression monitoring.

References

[1] Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M Mossi, and Amparo Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, 18(1):29, 2019.

[2] Bambang Krismono Triwijoyo, Widodo Budiharto, and Edi Abdurachman. The classification of hypertensive retinopathy using convolutional neural network. *Procedia Computer Science*, 116:166–173, 2017.

[3] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[4] Mikolajczyk A. and Grochowski M. Data augmentation for improving deep learning in image classification problem. *International Interdisciplinary PhD Workshop*, 2018.

[5] Jonghoon Jin, Aysegul Dundar, and Eugenio Culurciello. Robust convolutional neural networks under adversarial noise. 2016.

[6] Md Islam, Sheikh Asif Imran, Asiful Arefeen, Mahmudul Hasan, and Celia Shahnaz. Source and camera independent ophthalmic disease recognition from fundus image using neural network. 11 2019.

[7] Frid-Adara Maayan, Diamant Idit, Klang Eyal, Amitai Michal, Goldberger Jacob, and Greenspan Hayit. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Mar 2018.

[8] Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro Frangi. Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging*, PP:1–1, 03 2019.

[9] Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro Frangi. *Retinal Image Synthesis for Glaucoma Assessment Using DCGAN and VAE Models: 19th International Conference, Madrid, Spain, November 21–23, 2018, Proceedings, Part I*, pages 224–232. 11 2018.

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[12] Advanced guide to inception v3 on cloud tpu. `https://cloud.google.com/tpu/docs/inception-v3-advanced`. Accessed: 2021-01-11.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019.

[15] Lily Peng. Detecting diabetic eye disease with machine learning. `https://blog.google/technology/ai/detecting-diabetic-eye-disease-machine-learning/`.

[16] Lan Chang, Chen-Wei Pan, Kyoko Ohno-Matsui, Xiaoyu Lin, Gemmy C.M. Cheung, Gus Gazzard, Victor Koh, Haslina Hamzah, E. Shyong Tai, Su Chi Lim, Paul Mitchell, Terri L. Young, Tin Aung, Tien-Yin Wong, and Seang-Mei Saw. Myopia-related fundus changes in singapore adults with high myopia. *American Journal of Ophthalmology*, 155(6):991 – 999.e1, 2013.

[17] Ocular disease recognition using convolutional neural networks. `https://towardsdatascience.com/ocular-disease-recognition-using-convolutional-neura` 2020.

[18] Peking university international competition on ocular disease intelligent recognition (odir-2019). `https://odir2019.grand-challenge.org/introduction/`, 2019.

[19] Thomas G. Dietterich. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, page 1–15, Berlin, Heidelberg, 2000. Springer-Verlag.

[20] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[21] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.