# EYE DISEASE RECOGNITION USING FUNDUS IMAGES: A DEEP LEARNING APPROACH

**Pietro Obbiso**
University of Bologna

**Irene Pintos Castro**
University of Santiago de Compostela

## ABSTRACT

Deep learning models are widely used for classification of ophthalmic diseases using retinal fundus images. We use the ODIR-5K dataset to classify eye images, previously annotated with diagnostic keywords, using CNN models. From the available labels, we can satisfactorily (80-90%) classify between Normal and Cataract classes, but we fail (∼60%) to accurately classify other abnormalities as the diabetic retinopathy.

## 1 INTRODUCTION

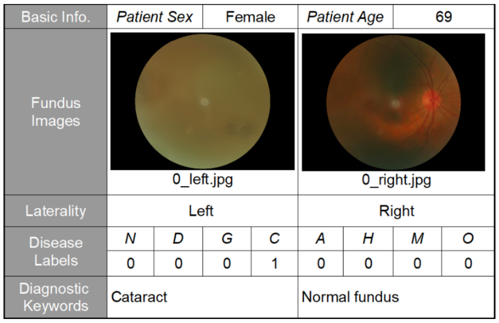
Deep learning (DL) is a state-of-the-art machine learning technique that has triggered an enormous interest in a wide variety of fields in the last few years. DL has shown to achieve significantly higher accuracies in natural language processing, computer vision and voice recognition, among others, compared with conventional techniques. In medicine and healthcare, DL has been applied to the analysis of medical imaging. DL systems have shown robust diagnostic performance in detecting various medical conditions, including tuberculosis from chest X-rays, malignant melanoma on skin photographs and lymph node metastases secondary to breast cancer from tissue sections.

Its application to ocular imaging, principally fundus photographs and optical coherence tomography (OCT), have been used for the detection of major ophthalmic diseases (e.g., diabetic retinopathy, glaucoma, age-related macular degeneration (AMD)). For example, (1) reports an accuracy of 88.5% in the detection of glaucoma abnormalities by extracting features automatically from the raw images by CNN and feeding them to a SVM classifier. See (2) for a summary of different DL systems and datasets recently used in ophthalmology.

In this project we are going to use a set of eye fundus images, previously annotated with diagnostic information, to apply a DL model for the classification of major eye diseases as diabetes, glaucoma, or cataract. As highlighted in (3), the most commonly utilized neural networks in state-of-the-art deep learning algorithms used in ophthalmic diagnosis with retinal fundus images are convolutional neural networks (CNN) and, thus, we will build our model using CNN.

## 2 DATA SET

ODIR-5K is a structured ophthalmic database collected by Shanggong Medical Technology Co. from different hospitals in China. This dataset includes color fundus images of left and right eyes from 5000 patients, along with basic information as the patient sex and age and diagnostic keywords from doctors. Such annotations are tagged to classify patients into eight labels: normal (N), diabetes (D), glaucoma (G), cataract (C), age-related macular degeneration (A), hypertension (H), myopia (M), and other diseases/abnormalities (O).



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Basic Info. | *Patient Sex* | | Female | | *Patient Age* | | 69 | |
| Fundus Images | 0_left.jpg | | | | 0_right.jpg | | | |
| Laterality | Left | | | | Right | | | |
| Disease Labels | *N* | *D* | *G* | *C* | *A* | *H* | *M* | *O* |
| | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Diagnostic Keywords | Cataract | | | | Normal fundus | | | |

Figure 1: The first structured ophthalmic record in ODIR-5K database.

It is worth noticing that one patient may contain one or multiple labels and these labels can refer to either left or right eye. From the 5000 patients, 3500 are included in the *Training sample* which we use as our dataset, since is the only sample that includes the diagnostic keywords. In Figure 1 we show an example of the structured information that is provided in the ODIR-5K dataset.

# 3 Methodology

## 3.1 Data Exploration

As a very first thing, we decided to treat the problem by not providing to the model as input both left and right eye images of a patient, as the dataset was created, but instead considering each eye image, left or right, as an independent image in such a way that we would be able to get predictions separately for each eye, to know for example which disease the eye has and how it should be treated. Thanks to the fact that the diagnostic keywords relate to a single eye, we were indeed able to create a new dataset simply by concatenating all the left-eye images with all the right-eye ones. In this way, each eye was assigned to a proper label.

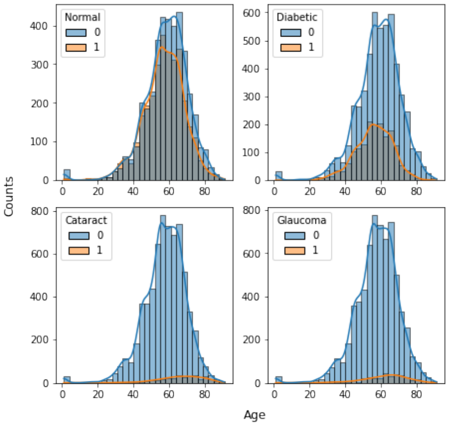


Figure 2: Age histograms of the first four classes (Normal, Diabetic, Cataract, and Glaucoma). Blue color is label=1, when the keyword diagnostic is present, and orange color is label=0, when the keyword is not found in the diagnostic.

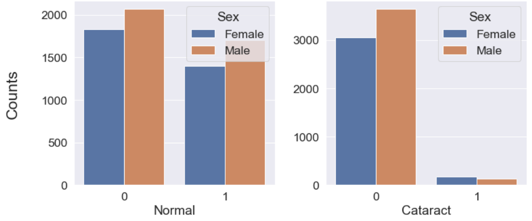


Figure 3: Histograms of the Normal and Cataract classes split by sex. Blue color is male while orange color represent female patients.

In Figure 2 we show the age histograms for the four classes that we classified based on the diagnostic keywords: Normal, Diabetic, Cataract, and Glaucoma. From the orange histograms, we see how the Normal label is the most common one, followed by the Diabetic label, and that the Cataract and Glaucoma classes are quite rare. Thus, from Figure 2, we are already seen that the dataset is unbalanced. Such inequality in the dataset classes is also visible in Figure 3, which shows the sex distribution of Normal and Cataract classes, that we will use for performing a binary classification using CNN.

The dataset in homogeneously sampled in age and balanced in sex, showing a thoughtful design. It is also clear that certain pathologies, as cataract is highly correlated with age. Unfortunately, due to the design of our algorithm, at this stage we are not going to include this extra information to train the CNN classifier.

## 3.2 Data pre-processing

Since the dataset was quite heavy and for the scope of this project we just needed some of the images contained in the whole dataset, we decided to create specific folders containing the images with the given disease we were interesting on. This approach was also useful for applying then the *Image Generator* function in order to generate our train and our test samples of images. This function takes the dataframe and the path to a directory and generates batches of image data with real-time data augmentation. This function allows us to easily rescale the images within the 0 to 1 range and resize them to 250x250 pixels (best value tested and the larger our hardware is able to process). As the main aim of the project is to focus on the classification using DL, we did not go deep in the image pre-processing to improve the classification (e.g., filtering, color scale).

## 3.3 Approach

For what concern the approach, as already mentioned in the Introduction section, we built our models using CNN. The idea was to create mini-batches, which consist in feeding to the model some portion of data, training and repeating with another portion. These portions are indeed called batches. The parameter *Batch size* defines how many examples will be extracted at each training step. After each step, the weights are updated. We selected batch size equal to 32, in order to avoid the overfitting problem. With small batch size, weights keep updating regularly and often, however, the downside of having a small batch size is that training takes much longer than with the bigger size.

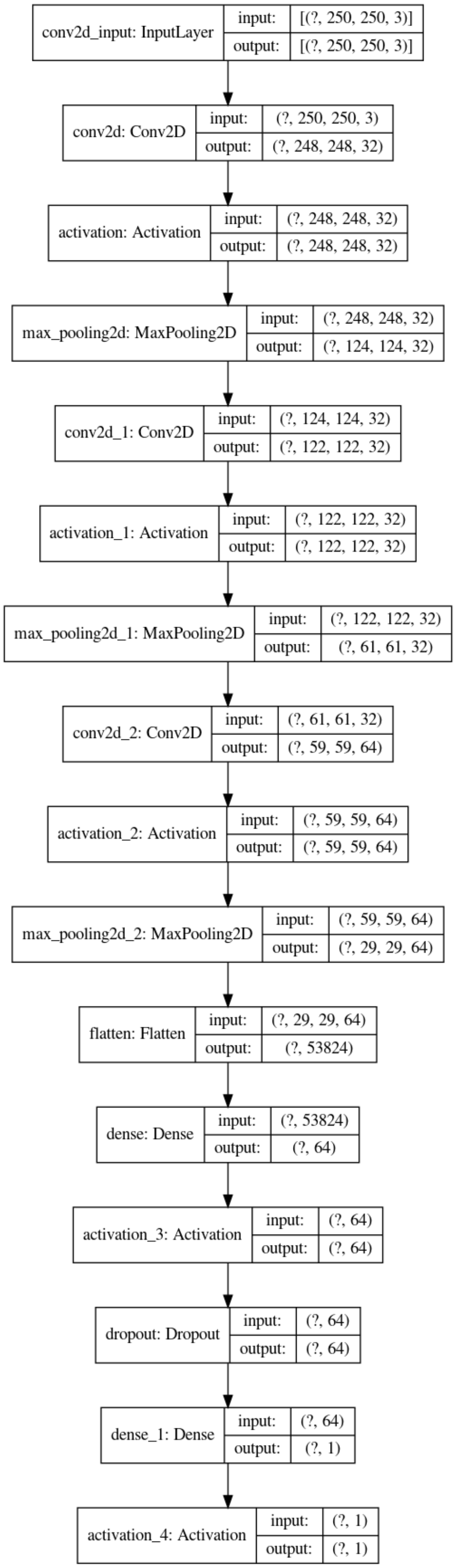


Figure 4: Architecture of the CNN model.

For what concern the architecture of the Neural Network itself, we decided to build it by starting with a first convolutional layer that takes as input 250x250 RGB images using a window of the size of 3x3 pixels to extract features and save them on a multidimensional array. In our case we set the number of filters for the first layer equals to 32.

After each convolution layer, a rectified linear activation function (ReLU) is applied. Activation has the authority to decide if neuron needs to be activated or not measuring the weighted sum. ReLU returns the value provided as input directly, or the value 0 if the input is 0 or less. To progressively reduce the spatial size of the input representation and minimize the number of parameters and computation in the network, a max-pooling layer is added. For each region represented by the filter of a specific size, in our example we used a filter of size (2, 2), it will take the max value of that region and create a new output matrix where each element is the max of the region in the original input.
Then a flatten layer is added. And finally, to avoid the overfitting problem, a dropout layer is added.
The last layer is a Dense layer with parameter equal to 1 since each image has only 1 output value, which is its label (0 or 1). Since we are facing a binary classification problem, the sigmoid activation function is applied to the last layer which converts each score to the final node between 0 to 1. Since we are using the sigmoid activation function, we must go with the binary cross-entropy loss. In our experiments we tried as activation function also the softmax function using as loss function the categorical cross-entropy. The entire architecture of our CNN is presented in the Figure 4.

At the end we obtain best results with the first combination, so we decided to build the neural network in that way. The selected optimizers were *Adam* and *rmsprop* and the chosen metric is *accuracy*.

## 4 RESULTS

In Figure 5 we show the values of accuracy and loss function obtained while training our CNN model for the balanced samples for each epoch. Though we used 15 epochs in most of our trials, here we show a 50 epochs model fit to demonstrate that, though the accuracy for the train sample keeps improving, the accuracy for the validation sample flattens after 15/20 epochs, showing that we have already reached the maximum accuracy (of ∼83%) for this model.

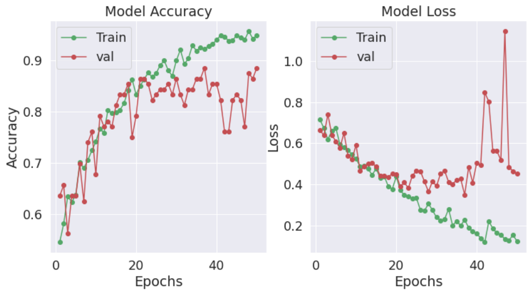


Figure 5: Accuracy and loss function history of the CNN model fit for the balanced training (green) and validation (red) samples.

The confusion matrix obtained as the result of the VGG19 model fit is shown in Figure 6. For that model we got an accuracy over the 85%. This relation between the true classes (on the y axis) and the predicted ones (on the x axis), show that this model returns very similar numbers for false positives and false negatives, no leaning the classification to either Normal or Cataract classes.

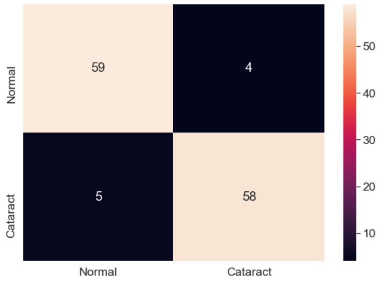
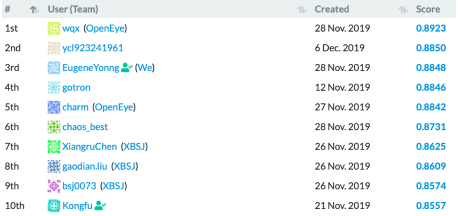


Figure 6: Consufion matrix corresponding to the VGG19 model fitting using the balanced samples.

After having experimented with the Cataract vs Normal scenario, we decided to try to implement the same model, maintaining the same architecture, with the Diabetic vs Normal scenario. We kept all the parameters resulted as the best in the previous scenario and considering as reference model the VGG19 one. However, the results were way lower than the the Cataract vs Normal ones. In particular, after 15 epochs, we obtained a value of 64.5% as accuracy.

One of the difficulties we have found to obtain a better trained model, arise from the size of the images. As mentioned before, we have to resize them to 150x150 and 250x250 pixels to be able to run the models, or we run out of memory space. The original size of the images, though highly variable, is of the order of 1800-1200 pixels. So, when we resize them to 1/10 or even more of their original size, we are irredeemably loosing information. Such information, in the small details, is fundamental to properly identify the signs of, for example, diabetic retinopathy or glaucoma.



| # | User (Team) | Created | Score |
|---|---|---|---|
| 1st | wqx (OpenEye) | 28 Nov. 2019 | 0.8923 |
| 2nd | ycl923241961 | 6 Dec. 2019 | 0.8850 |
| 3rd | EugeneYonng (We) | 28 Nov. 2019 | 0.8848 |
| 4th | gotron | 12 Nov. 2019 | 0.8846 |
| 5th | charm (OpenEye) | 27 Nov. 2019 | 0.8842 |
| 6th | chaos_best | 28 Nov. 2019 | 0.8731 |
| 7th | XiangruChen (XBSJ) | 26 Nov. 2019 | 0.8625 |
| 8th | gaodian.liu (XBSJ) | 26 Nov. 2019 | 0.8609 |
| 9th | bsj0073 (XBSJ) | 26 Nov. 2019 | 0.8574 |
| 10th | Kongfu | 21 Nov. 2019 | 0.8557 |

Figure 7: Current leader-board of the *Grand Challenge*.

The code that we have developed for completing this project, including the CNN model fitting, is available at `https://www.kaggle.com/irenepintoscastro/notebook-fmlcv-project`

## 5 Conclusions and future works

We have been able to successfully classify eye fundus images with cataract from the healthy ones using deep learning CNN models. However, the classification of other diseases as the retinophaty associated with diabetes, or even a multi-classification including also glaucoma, require further training. Detecting myopia or cataract is a much easier task because these images vary a lot from each other and from the normal fundus, while the individual classification of other pathologies, which require more details to be detected, and, consequently, the classification of multiple categories, are a hard challenge as we can conclude also from the current leader-board of the challenge competition, where the highest accuracy score is 89%.

Regarding future works, several are the further analysis that one could make. First of all, with the possibility in the future of having more data classified as certain disease, it would not be necessary to perform any augmentations, as sufficiently enough image variations would be provided. Moreover, by having a more efficient computational power, it would be possible to increase the dimension of the images obtaining a better resolution and as a consequence better results in terms of accuracy. Then, one could think about modifying the architecture of the neural network, either removing or adding different layers, and at the same time playing with all the parameters involved (e.g. activation function, loss function, optimizer and so on...). Eventually, we only considered the cases of Diabetic vs Normal and Cataract vs Normal. However other combination could be made, for example by considering other diseases vs Normal, or adding to the dataset we used images of another class or even more challenging considering the entire

ODIR dataset transforming the problem into a multi-label classification one.

## References

[1] B. Al-Bander, W. Al-Nuaimy, M. A. Al-Taee, and Y. Zheng, "Automated glaucoma diagnosis using deep learning approach", in *2017 14th International Multi-Conference on Systems, Signals Devices (SSD)*, 2017, pp. 207–210.

[2] Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong, "Artificial intelligence and deep learning in ophthalmology", *British Journal of Ophthalmology*, vol. 103, no. 2, pp. 167–175, 2019.

[3] Sourya Sengupta, Amitojdeep Singh, Henry A. Leopold, Tanmay Gulati, and Vasudevan Lakshminarayanan, "Ophthalmic diagnosis using deep learning with fundus images – a critical review", *Artificial Intelligence in Medicine*, vol. 102, pp. 101758, 2020.