Deep Learning Application in Primary Open-angle Glaucoma Detection, Prediction, and Analysis

Eric Zhewen Li*
zl733@cornell.edu
Jacobs Technion-Cornell Institute,
Cornell Tech, Cornell University
New York, NY, USA

Flora Yu Shen*
ys2223@cornell.edu
Jacobs Technion-Cornell Institute,
Cornell Tech, Cornell University
New York, NY, USA

ABSTRACT

Primary open-angle glaucoma (POAG) is one of the leading causes of blindness for people over the age of 60. It is expensive and time-consuming to manually detect POAG from fundus photographs. Last semester, we implemented an automated model using the Deep Learning approach to diagnose whether an eye has POAG or not given its current fundus image and achieved satisfactory results. Continuing from last semester's work, we now focus on automated models to predict POAG onset ahead of time. With the new problem statement, we have carried out more robust data splitting and labeling strategies, implemented novel data pre-processing techniques, and exploited different Machine Learning models, including several variations of the Convoluted Neural Network as well as the Vision Transformer. With this new setup, we are able to achieve a successful prediction model, with an area under the receiver operating characteristic curve of 0.763.

1 INTRODUCTION AND PROBLEM DEFINITION

Many forms of primary open-angle glaucoma (POAG) have no warning signs. Hence, it is important to diagnose its early stages so that patients can be treated appropriately and help them get health back. With the fast development of deep learning technologies and their applications in the bio-image analysis field, we can train good prediction models by learning from a massive amount of image data. With such power of neural networks, we would like to further utilize this technology to detect abnormality from fundus images, give POAG diagnosis in real-time, and predict the development of the disease ahead of time, thus preventing many tragedies from happening.

Last semester, we focused on the real-time detection of POAG from a single eye fundus image. After training our machine learning model on the fundus image dataset, our model would decide whether the patient currently has POAG given the fundus image. The model was able to achieve an accuracy of 90.94%, an area under the receiver operating characteristic curve (AUC) of 0.876, and F-score of 0.4497.

This semester, we pivoted to the preventative prediction of POAG. Given fundus images of one visit, our model would decide on whether the patient will develop POAG in x-year time, where $x \in \{1, 2, 3, 5, 7\}$. We converted our prediction problem into a classification problem, with input as a single image or pair images taken at each time visited, and output as a binary label of if the patient will develop the disease within x-year time. In our study, we used Convoluted Neural Network (CNN) and Visual Transformer for Image Classification (ViT) from hugging face as the main backbone models. We have also experimented with several variations of CNNs pre-trained on the *ImageNet* dataset, including VGG-16, ResNet-152, and DenseNet-201.

Due to the severe dataset imbalance problem, our primary work in this study uses data pre-processing techniques, such as image augmentation, class weight penalty, different data labeling methods, and data discarding, to ease the imbalance problem.

2 RELATED WORK

During the literature review, we found a similar paper [6]. Their paper assesses the accuracy of deep learning models to predict glaucoma development from fundus photographs several years before disease onset. They proposed an approach to train deep learning models on 85% of the OHTS data (training set), and validate them on the remaining 15% held-out set (validation set). This research shows that deep learning models can predict glaucoma development prior to disease onset with reasonable accuracy. When predicting glaucoma development 4-7 years before onsite, this research achieved an AUC of 0.77 (95% confidence interval 0.75, 0.79).

When we set up our experiment, we also tried to replicate their work as a reference. However, we realized that their work pre-selected patient samples: negative samples all came from patients who never had POAG onset throughout the trial, while positive samples all came from patients who had POAG developed during the trial. This setting converted the whole problem to detect if the sample is from positive or negative patients, which lowered the difficulty of the problem.

3 DATASET

We used the same dataset as last semester. The fundus photographs of this study (i.e., the dataset) come from the Ocular Hypertension Treatment Study (OHTS) [1]. This dataset consists of 35,092 eye fundus images, throughout a 15-year clinical study from 1,636 patients, where 279 patients have developed POAG during the trial (either left-, right-eye, or both eyes) and the remaining 1,357 patients have not developed POAG throughout the study.

3.1 Data Pre-Processing

There are three types of images in the dataset, $-L$, $-R$ and $-S$,. $-L$ and $-R$ refer to images taken from the left- and right-angle of the eye during fundus imaging for a single visit, while $-S$ images include both angles onto the same image via a single simultaneous shot of the eye. To build consistency in our dataset, we developed a

*Both authors contributed equally to this research.

script to cut the $-S$ images by the vertical middle line and save both sides as $-SR$, $-SL$ respectively. Since we wanted to keep the central retina part of the image, we cut out the peripheral black border. For $-L/R$ images, the original image size is 1502×1000 pixels, and $-S$ images have an original image size of 751×1000 pixels. In order to keep the area in the image corresponding to roughly where the central retina is, we pre-processed the images by cutting from its geometry center with a square of size length 350 (for $-L/R$ images) and 250 pixels (for $-S$ images), resulting in final image sizes of 700×700 and 500×500 pixels, respectively.

3.2 Data Splitting

To train our machine learning model, we need to split our dataset into training, development, and testing sets. Due to the highly imbalanced nature of our experiment data, we have implemented new splitting methods this semester to ensure our training process is fairer and hence better models after training. In addition, we have also implemented 5-fold cross-validation in our experiment settings for the same reason.

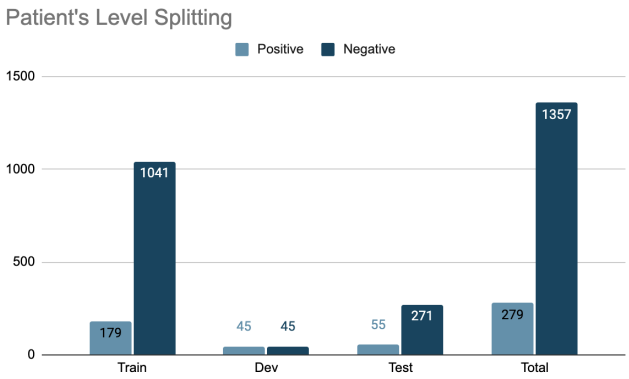

Figure 1: Dataset split distribution on patients level

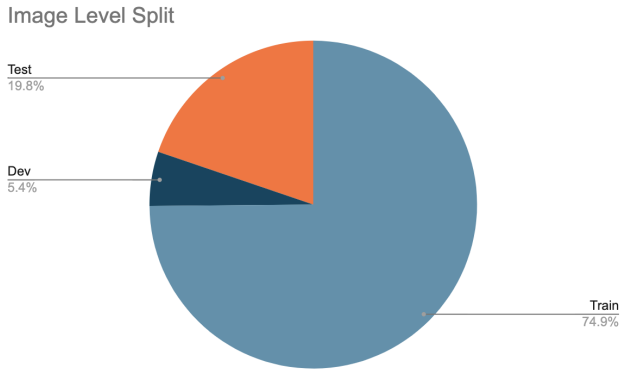

Figure 2: Dataset split distribution on the image level

We implemented a patient-level split as we want our model to see the entire course of disease for one certain patient. Hence we divided each category of patients (POAG vs. non-POAG) into five groups of equal number - this helps us to implement a 5-fold cross validation and validate the robustness of our experiments. Throughout our project for this semester, we did not implement the cross validation due to time constraints. We used all the images from the first group as our testing set and those from the remaining group (second, third, fourth, and fifth) as our training and validation sets. With this new splitting method, in each fold of the dataset, we have 179 POAG patients and 1,041 non-POAG patients in the training set, 45 POAG patients and 45 non-POAG patients in the validation set, and 55 POAG patients and 271 non-POAG patients in the test set. With this patient-level split, our final dataset consists of 26271 training images, 1878 validation images, and 6943 testing images, which is roughly $75\% - 5\% - 20\%$ ratio.

3.3 Data Transformations

Since all our selected models require input images of size $(224, 224, 3)$, representing their width (in pixels), height (in pixels) and RGB color channels, we implemented the following image transformations for our development and test sets before we feed them to the network.

- Resize to 256×256
- Center crop to 224×224
- Standard *ImageNet* normalization

Since images in our dataset may have different sizes, we resized them to 256×256. In order to get the center part of our fundus image, we then cropped the center 224×224 of each image. Then we applied a normalization on each channel of the input image.

3.4 Data Augmentation

To prevent our models from overfitting during training, we performed the following image augmentations and transformations on the training dataset before feeding them into the network and starting training.

- Random resized crop to 224×224
- Random horizontal flip
- Random vertical flip
- Random rotation from -30 to 30 degrees
- Random contrast, saturation, and hue changes, controlled by `ColorJitter` parameters
- Standard *ImageNet* normalization

In the normalization step, all of our datasets (train, development, and test) are normalized to have *mean* $=[0.485, 0.456, 0.406]$, and *standard deviation* $=[0.229, 0.224, 0.225]$, each value corresponding to the RGB channel, as expected by our pretrained models.

3.5 Data labeling

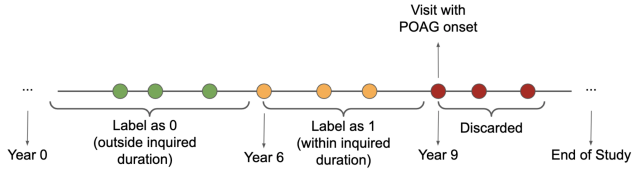

Figure 3: An example timeline for POAG patients' data labeling, with a 3-year inquired duration

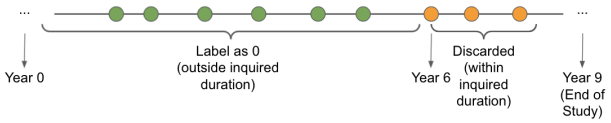

Figure 4: An example timeline for non-POAG patients' data labeling, with a 3-year inquired duration

In order for our model to make predictions instead of detections, we need to re-label each image in the dataset as well. We define patient's POAG onset year to be `round(onset date / 365)`, where it rounds up if the onset date is less than 182 days past the beginning of any calendar year and rounds down if that is at or more than 182 days after the beginning of any calendar year. Then we define the patient's course of disease to be of range [onset year × 365, onset date).

We label all the images graded between the patient's course of disease of label 1, as within x-year after this fundus image was taken, this patient has developed POAG. We label all the images before the course of disease of label 0, and discard all the images taken on and after this patient's onset date. Using our baseline experiment settings of $x = 3$, we can label 25629 negative and 642 positive images in the training set, 1693 negative and 185 positive images in the validation set, and 6716 negative and 227 positive images in the testing set.

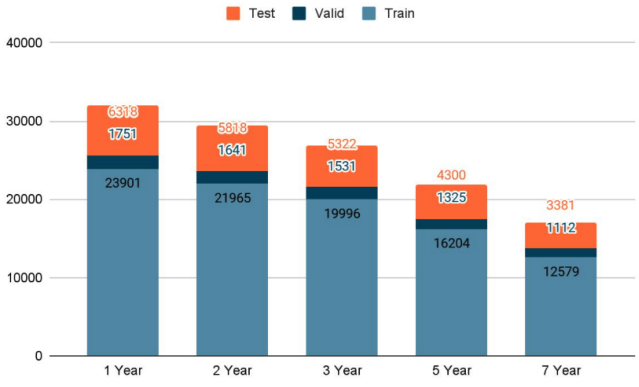

Figure 5: Dataset distribution on trace-back years labeling

3.6 Image Pairing

There can be multiple images taken on one visit. Thus we also tried to pair left eye and right eye images taken on the same visit, and merge two three-channel images into one single six-channel image. For a visit with only one image taken, we duplicated the single image and merged the two identical images to six-channel. The purpose is to see if feeding in more information for one visit would affect the overall model performance.

4 METHOD

For this project, we are working on the image recognition problem; hence, we are able to use some pre-trained image recognition models as base models before we fine-tune them to fit on our dataset. We have experimented with different pre-trained models, transfer learning techniques, and hyperparameters, to benchmark their performances on predicting POAG onset from fundus photos.

4.1 Convoluted Neural Network

Since we converted our prediction problem to a classification problem, we used a set of methods similar to the detection method. We kept using Convolution Neural Network (CNN) and transfer learning CNN models trained on *ImageNet*.

Since we converted our prediction problem to a classification problem, we used a set of methods similar to the detection method. We used Convolution Neural Network (CNN) and transfer learning CNN models trained on *ImageNet*. We used VGG-16 [5], ResNet-152 [3] and DenseNet-201 [4].

4.2 Vision Transformer

We also explored Vision Transformer this semester. Transformers are the most popular method for Natural Language Processing (NLP) tasks. The work on Vision Transformer (ViT) proposes a transformer model on image sequences for the classification task. It demonstrates strong performance that beats most state-of-the-art CNN networks on various image recognition datasets using lesser computational power. We think using the image transformer can be a good learning practice for us. Therefore, we chose to adapt ViT to our own image classification task. For implementation, we used the Hugging Face API. The model was proposed in a recent paper [2]. The vision transformer has an encoder and decoder structure, and was pre-trained using a resolution of 224×224 on *ImageNet*. In order to feed into the encoder, each image is split into a sequence of fixed-size non-overlapping patches, which are then linearly embedded. A [CLS] token is added to represent an entire image, which can be used for classification.

5 EXPERIMENTS

5.1 Implementation Details

We implemented our experiments using Python and PyTorch libraries. All the training is conducted on the Scientific Computing Unit (SCU) provided by the Weill Cornell Medical College.

5.2 Experiment Results

During the training process, we found that our model may overfit on the training dataset, where the training loss continues to decrease but the validation loss no longer decreases anymore, as shown below in Figure 6. To compare our model training results fairly, we have implemented the early stopping techniques, where we would have stopped the training process when validation loss no longer decreases.

After fine-tuning all the hyper-parameters and implementing the early stopping technique, we then benchmarked with different base models outlined earlier in Section 4, using learning rate = 1×10^{-6} and color jitter = 0.1. The results are shown below in Table

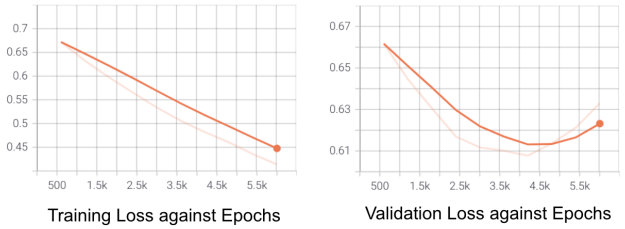

Figure 6: Example training and validation loss to show the importance of early stopping

1. We can see that DensetNet-201 performs the best among all the CNNs - we think this makes sense as it has the most layers to train and hence can detect more delicate signals in our dataset.

Table 1: Base model benchmarking, with 3-year inquired range, 1e-6 learning rate, and 0.1 color jitter

Base model	Early stopped at	Test AUC
VGG-16	4	0.737
ResNet-152	2	0.734
DenseNet-201	**5**	**0.761**
ViT	9	0.595

It is also surprising to see that ViT performs way worse when compared to CNNs. We think this is due to the fundamental infrastructure difference between ViT and CNN and we should fine-tune our hyper-parameter set for ViT separately to achieve the best performance.

Additionally, we have also benchmarked on the inquired ranges. The nature of POAG prognosis is very slow and hence we have not included the 1- and 2-year inquired ranges in our benchmarking. Among those we tested, as shown below in Table 2, we found that 7-year inquired range has the best performance while 3-year has a very close performance as well.

Table 2: Inquired range benchmarking, with DenseNet-201 base model, 1e-6 learning rate, and 0.1 color jitter

Inquired range	Test AUC
3-year	0.761
5-year	0.757
7-year	**0.763**

Lastly, we compared our model performances by training with one image and two images from each visit, as outlined above in Section 3.1. The test results are shown below in Table 3. These results show that there is no major difference between the two approaches but the model using only one randomly selected image from each visit has a higher performance than that using both images from each visit. We think in the later setup, the machine learning model sees duplicate images from the same visit and can be easily confused and converge to detecting minor differences on the fundus images that are not related to the POAG prognosis (e.g., exposure settings)

Table 3: Training technique comparison, with 3-year inquired range, 1e-6 learning rate, 0.1 color jitter, and DenseNet-201 base model

	Negative	Positive	Test AUC
One image	**25,698**	**1,151**	**0.761**
Two images	51,396	2,302	0.742

5.3 Additional Benchmarking Results

We have also benchmarked our model training results with respect to learning rate and color jitter parameters. As shown in Tables 4 and 5 below, all other sets of hyper-parameters would result in models overfitting on the dataset, and a decrease in the test AUCs.

Table 4: Learning rate fine-tuning, with 3-year inquired range and DenseNet-201 base model

Learning rate	Test AUC
1×10^{-3}	0.5855
1×10^{-4}	0.5948
1×10^{-5}	0.6154
$\mathbf{1 \times 10^{-6}}$	**0.7032**
1×10^{-7}	0.6758

Table 5: Color jitter fine-tuning, with 3-year inquired range, DenseNet-201 base model, and 1e-6 learning rate

Color jitter	Test AUC
0	0.7032
0.1	**0.7668**
0.2	0.7583
0.3	0.7432
0.5	0.7547

6 CONCLUSION AND DISCUSSIONS

In this project, we applied deep learning models to not only detect POAG in real-time but also predict the disease's onset time several years ahead of time from eye fundus photographs. Our extensive experiments show very promising results in all three problems of diagnosis, short- term, and long-term prognosis. This shows that machine learning models can detect the delicate signals present in eye fundus photos and predict POAG onset several years ahead of time. This could help patients detect POAG in much earlier stages and improve their treatment outcomes, and eventually generate a huge impact in the Health Tech field.

When compared with our baseline paper mentioned in Section 2, as shown below in Table 6, our experiments result in a very close performance in all three AUCs. Although our test AUCs are a little lower, we believe our experiment setup is more robust because the baseline paper has manually sanitized the dataset by discarding 24% of the bad images while in our setup, the machine learning models can still generalize with additional noise exist in the training set. Additionally, we have included all the initial negative images from patients that developed POAG in later stages and our problem definition is more aligned with the final goal to predict POAG onset ahead of time, rather than predict whether the image comes from a POAG or non-POAG patient.

Table 6: Training technique comparison, with 3-year inquired range, 1e-6 learning rate, 0.1 color jitter, and DenseNet-201 base model

Experiment	Best diagnosis test AUC	Best 3-year prognosis test AUC	Best 7-year prognosis test AUC
Baseline paper	0.945	0.880	0.765
Our setup	0.876	0.761	0.763

REFERENCES

[1] 2007. Validated Prediction Model for the Development of Primary Open-Angle Glaucoma in Individuals with Ocular Hypertension. *Ophthalmology* 114, 1 (Jan. 2007), 10–19.e2. https://doi.org/10.1016/j.ophtha.2006.08.031

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[4] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. *CoRR* abs/1608.06993 (2016). arXiv:1608.06993 http://arxiv.org/abs/1608.06993

[5] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[6] Anshul Thakur, Michael Goldbaum, and Siamak Yousefi. 2019. Predicting glaucoma prior to its onset using deep learning. *bioRxiv* (2019). https://doi.org/10.1101/828681 arXiv:https://www.biorxiv.org/content/early/2019/11/02/828681.full.pdf