Shiping Li

DABETIC RETINOPATHY DETECTION

1. INTRODUCTION

a) Background

High blood pressure not only can cause heart and kidney problems but also eye disease. One of the common eye diseases is diabetic retinopathy, which can be found in many working-aged adults. The early stage of diabetic retinopathy usually does not have any symptoms but can develop severe vision loss and even blindness. Currently, detecting diabetic retinopathy is a time-consuming process where patients' fundus images are examined and diagnosed by highly trained clinicians manually. Arvind Eye Hospital in India expressed its concern about the eye health of people in rural areas. People living in rural areas have a higher risk of developing blindness because fundus screening is difficult to conduct under poor medical conditions. Even if screening is conduct successfully, the screen images are required professional doctors to review one by on. Such that, the whole diagnosis process becomes difficult and inefficient. To avoid more people missing the optimal treatment time, Aravind Eye Hospital in India captured screen images from rural areas, and hoped to leverage technology to screen images automatically and gain information on the severity of the condition.

b) Objective

The objective of this project is to provide a way of accelerating patients' diagnosis time and ensuring the effectiveness of the diagnosis results through deep learning. Therefore, we will end up with a model that is not only able to identify diabetic retinopathy from healthy eyes, but also grading severity of the disease. In the future, we can develop a model to detect other sorts of diseases like glaucoma and macular degeneration.

2. DATASET INFORMATION

The data are the fundus images captured from the rural areas by Aravind Eye Hospital. The total number of the images is 5590. Among them, 3662 images are the training, each of which comes with a unique id number and is rated on a scale of 0 to 4 by a clinician, based on the severity of diabetic retinopathy. The rest images are test data.

As mentioned above, the severity of diabetic retinopathy can be divided in to 5 categories or 5 stages, where

- stage 0 – No DR. No DR indicates healthy eyes.

- stage 1 – Mild. The earliest stage of diabetic retinopathy is often where swelling begins in retina's vessels.
- stage 2 – Moderate. In this stage, increased swelling of microscopic blood vessels begins to block blood flow to the retina, preventing proper nourishment.
- stage 3 – Severe. A larger section of blood vessels in the retina become blocked, resulting in insufficient blood flow to this area. The body receives signals to start growing new blood vessels in the retina.
- stage 4 - Proliferative diabetic retinopathy. This is an advanced stage of the disease in which new blood vessels grow in the retina, causing blurriness, a limited field of vision, and even blindness [1].

Since our goal is to detect the if a fundus image has diabetic retinopathy and its possible condition, our target value is exactly the categorical five stages of the disease.

3. DATA EXPLORATORY ANALYSIS

a) Distribution of the Target Value

The count plot shows that the label class of severity is distributed unequally. No DR accounts for nearly half of the training data, and Moderate accounts for around 27%, adding up to 77% approximately. At the same time, Severe and Proliferative account for less than 15% in total. The problem here is that the machine learning will be more biased to the majorities especially No DR. People with diabetic retinopathy will be wrongly informed of no disease, which is the worst situation and completely deviates from our intention. This will lead to increasingly dangerous disease condition that threaten people's eye function and life. Thus, we will address this imbalance problem in the later section. The approaches will involve in choosing the right loss and metrics.

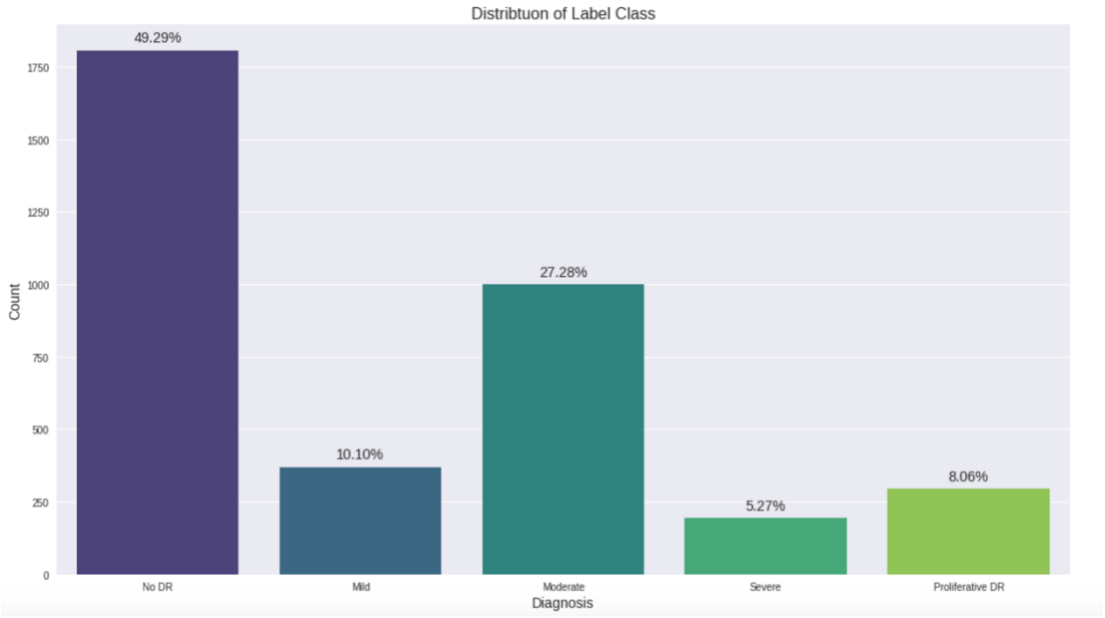


figure 1: Distribution of Label Class

b) Samples of fundus images across different severity

The following images (figure 2) are random sample of 5 different stages of diabetic retinopathy. As the severity level increases, the fundus tend to be more cloudy and spotty. Severe and Proliferative DR have stronger signs of blood clots and growth of blood vessels. Besides, the sizes of the images are varying, so they need to be adjusted to the same size before passing into a deep learning neural network. Also, some of the images have larger black region than the others, which may disturb the model from learning the correct information. We may crop the black region if necessary, but since the area of the black region seems to be random and not correlated to the stages, they can just be treated as random noises and lower generalization error. Another thing we were concerned is some images have lower quality such as low brightness, saturation which may increase the difficulty of the model learning the details. Next, we will explore the possible way to detect the low-quality images and enhance their quality.

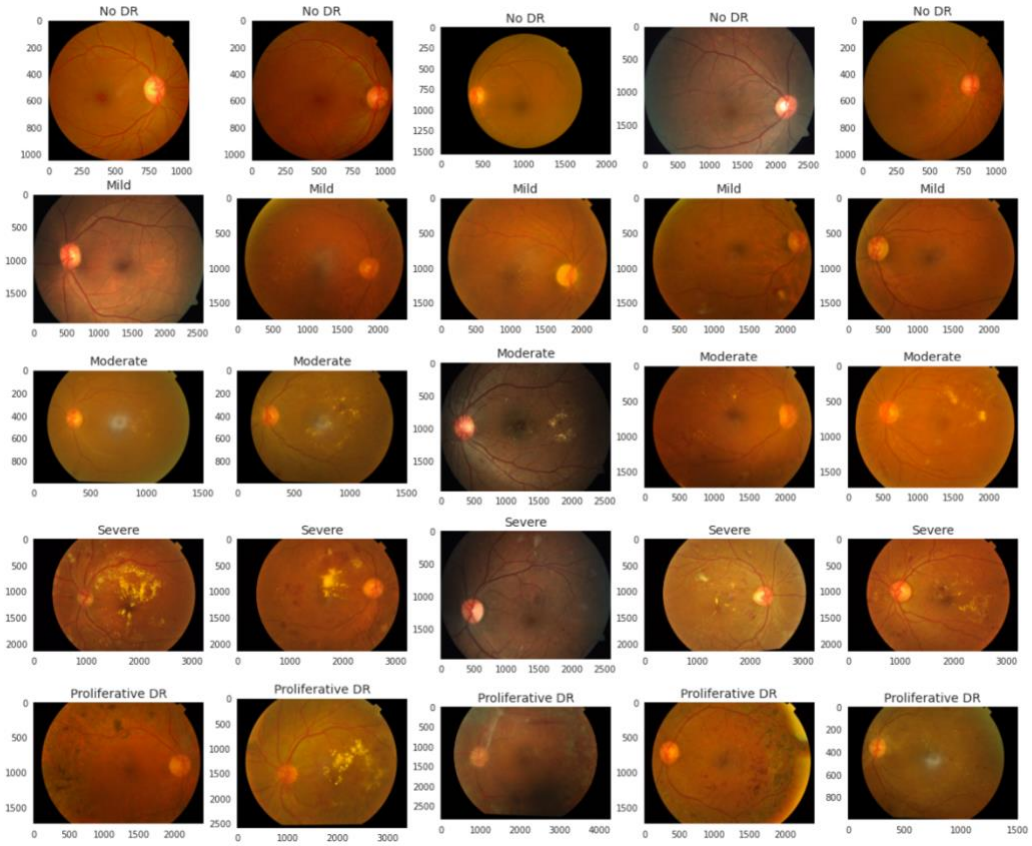
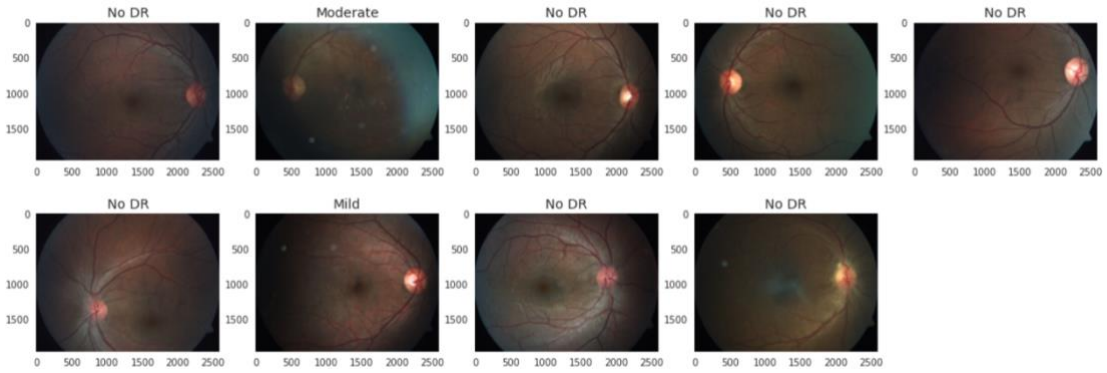


Figure 2: Samples of 5 different stages

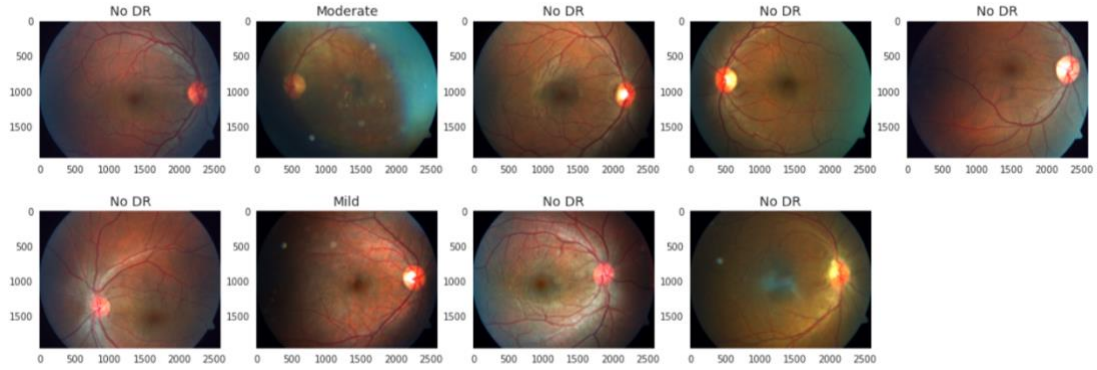
c) Detect low-quality images and enhance image

Instead of representing colors by red, green, and blue of an image, we represented them by hue, saturation, and value where hue carries color information, while saturation and value carry greyness and brightness information respectively. For the purpose of exploration and reducing memory, We sampled 200 images for testing, converted them from RGB to HSV color, and obtained their mean value of saturation and brightness. When the mean values of saturation

and brightness of an image were under the thresholds, then it was defined as a low-quality image. Here, we were only interested in saturation and brightness because all the fundus images are in their original color (yellow), so their hue values are expected to be close, and adjusting their hue values should not provide any additional information. After identifying which images with low saturation and brightness, we enhance their quality by scaling the saturation and brightness by 1.4 and 1.6, respectively. The following images are low-quality images identified in the sample of 200 images and their modified versions. (Note: These modified images were not used in our ultimate model training. This exploration aims to provide some insights for the research or develop the model in the future.)



(a) Original Images



(b) Improved images

Figure 4: A Sample of improved images

4. MODEL TRAINING AND TUNING

a) Some preparations

In the whole process of model training, we will leverage the pre-trained model of Residual Neural Network 50, namely ResNet50. ResNet50 has the advantages of increasing the network depth while using reducing the complexity of the model with fewer parameters. It achieves better performance compared to the traditional CNN models for image classification problems yet accelerates the speed of training deep networks at the same time.

The metric to evaluate the model performance is Quadratic Weighted Kappa or QWK. The reason why we use QWK is QWK takes the similarity of the predicted rating and actual rating into account. The 5 stages of retinopathy are hierarchical, so the gap between the predicted and the actual rating is meaningful. When the predicted rating is not the same as the actual class, we still wish to give higher credits to those predictions closer to the actual rating, compared to those further from the actual rating. For example, predicting severe DR for a proliferative fundus image gets a higher QWK score than predicting no DR. QWK takes values from 1 to -1. A perfect score of 1 is granted as the predicted ratings and actual ratings are exactly the same while the least score of -1 is given as the predicted ratings are furthest away from the actual ratings. A score of 0 means the performance of a model is close to random prediction. The formula of QWK is shown as below.

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}.$$

- N : the number of categories
- $W_{i,j}$: the weight of predicting j when actual rating is i
- $O_{i,j}$: the counts of i rating (actual) that were predicted as j rating
- $E_{i,j}$: the expected counts of rating i rating (actual) that are predicted as j rating.
- K : the QWK score

Utilized the image data generator to split data into training set with 2930 images, a validation set with 732 images, and a test set with 1928 images. Then rescaled the images by 1/255 and resize them to 224 x 224.

b) Warmup Training

In the warmup phase, the model started training with a small learning rate that grows linearly at the end of each iteration. The total number of warmup epochs was 5 and the start learning rate was $1/5\text{e}^{-3}$. We increased the learning rate by $1/5e^{-3}$ each time and the learning rate would ultimately reached 1e^{-3} at the last epoch. The reason behind the warmup is to avoid the model weights fluctuating drastically and deviating from the optima in early training. In addition to that, I set the top 5 layers trainable and the rest frozen to maintain the stability of the warmup model and shortened the training time.

In order to reduce the training time, we tracked the execution time of warmup training with batch sizes 16, 32, 64, respectively. The results showed no significant difference among them. Execution time of batch size 16 spent 2 more minutes approximately within 5 epochs, compared to the others. The QWK scores were close to 0 over time for all of them, so either batch size of 64 or 32 should be a reasonable choice. But since smaller batch size tends to

achieve the better generation performance, so our final choice was batch size 32. Intuitively, a model learns by seeing less data each step but it gets more chances to be exposed to new data through more training steps each iteration.

c) Select the optimal learning rate and fine tune the complete model

As soon as our warmup stage completed, it was time to find the optimal learning rate for our model with all layers trainable.

When the network's loss decreases the fastest in a certain range of learning rates, we could select a learning rate within the range. Then we can performed a simple experiment where learning rate increased on an exponential scale each mini batch, where the learning rate started with $1e^{-5}$ and ended with $1e^{-2}$. At the same time, we can track the loss at each increment. Our results showed that the loss dropped most rapidly with the learning rates in the blue region (figure 5).

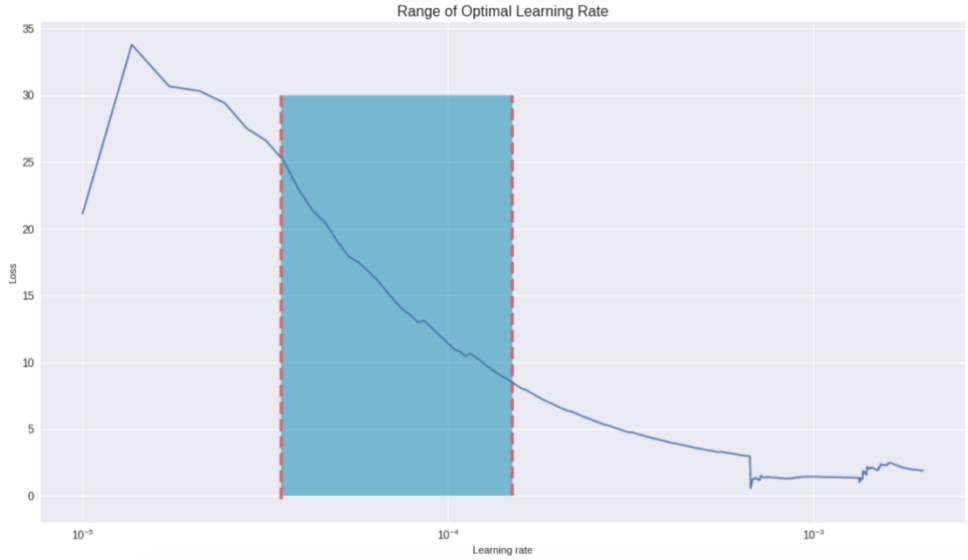
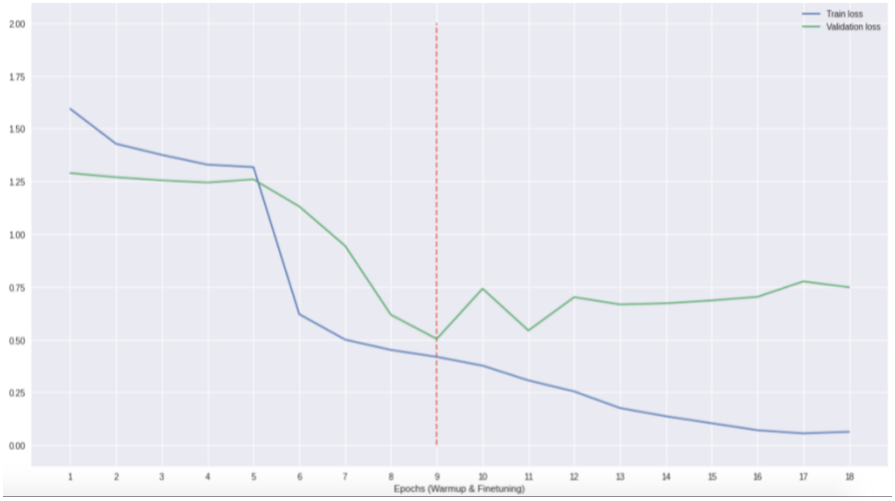


Figure 5: Optimal Range of Learning Rates

Ultimately, I selected $6e^{-5}$ as the start learning rate for fine tuning our complete model, of which all layers were trainable. In this phase, I set the total number of epochs to be 30, utilized validation loss as the monitor in early stopping, decayed the learning rate by 0.5 if validation loss did not improve in 3 epochs, and saved models when the best validation loss and QWK score reached the highest.

Figure 6(a) and (b) shows that the validation loss achieves the best at the 9th epoch and the corresponding QWK score reaches its first peak after a rapid climb. The gap between train loss and validation loss is very small. After the 9th epoch, they start to diverge, which means the model start learning the noises of the training set. Now, we obtain our first candidate model at the 9^{th} epoch.

(a) Train Loss & Validation Loss



(b) QWK score On the Validation Set

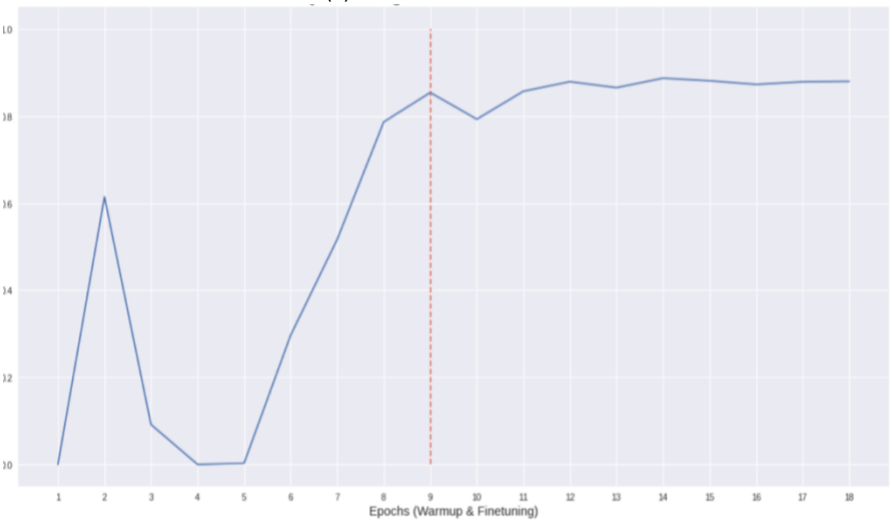


Figure 6: Fine Tune the Complete Model

For the second candidate model, I followed the whole training process of the first candidate model, from warmup training, optimizing the learning rate to finetuning. The only difference between the first and the second candidate models was the loss function. The first candidate model applied the **average cross-entropy loss function** whereas the second candidate model applied the **weighted cross-entropy loss function.** Our dataset was imbalanced as mentioned above, so I was interested in if adding more weights to minorities would lead to better results.

4. MODEL EVALUATION

a) Candidate Model 1(average cross-entropy loss function)

1. *Confusion Matrices*
   - The patterns on the training and validation confusion matrices were very close, which indicates the model performs consistently on the unseen data.

- The model was good at identifying No DR and Moderate classes but can do better for the rest of the classes in future improvement. The recalls of the validation set achieved 96% on No DR while 35% and 46% on Severe and Proliferative DR. The difference in performance was caused by unequally distributed classes.
- Severe, Proliferative, and Mild DR were mostly misclassified as Moderate.
- In addition to Moderate, the model had a hard time distinguishing Severe from Proliferative DR. There was a 35% chance a Severe sample classified as Moderate and 25% chance classified as Proliferative DR.

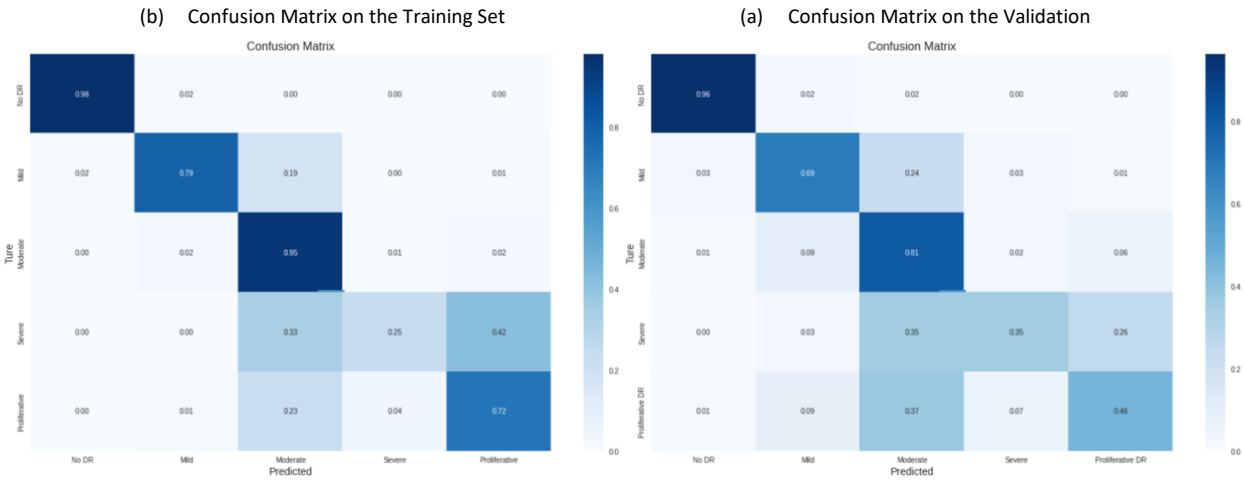


Figure 7: Confusion Matrices of Candidate Model 1

2. *Quadratic Weighted Kappa Scores*
    - On the training set: 0.938
    - On the Validation set: 0.854

    Just remind that perfect prediction yield a score of 1. We have 0.938 and 0.854 respectively, which are very decent scores.

b) *Candidate Model 2(weighted cross-entropy loss function)*

1. *Confusion Matrices*
    - The model seems to overfit the training data. There was a notable difference between patterns of the training and validation confusion matrices. The model performed perfectly on all classes in training data but not in validation.
    - Compared to the last model, the recalls on minorities were a bit better, given comparatively higher recalls on Severe and Proliferative.

a) Confusion Matrix on the Training Set

b) Confusion Matrix on the Validation Set

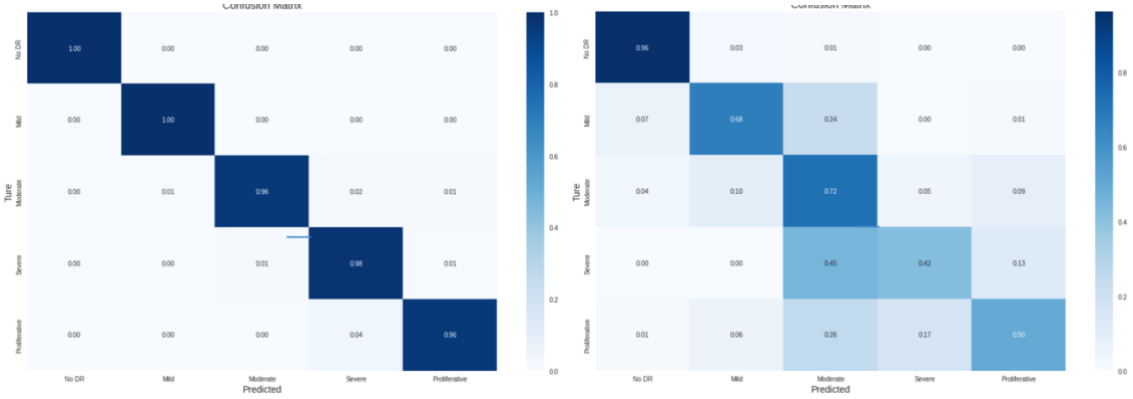


Figure 8: Confusion Matrices of Candidate Model 2

2. *Quadratic Weighted Kappa Scores*
    - On the training set: 0.992
    - On the Validation set: 0.862

    Again, a large gap presented between training and validation QWK scores. This reveals that the model tried to fit the noises of the training data.

5. FINNAL MODEL AND PREDCTION ON THE TEST SET

Even though the QWK scores of this model were slightly higher than 0.938 and 0.854 of the first candidate model, we chose the first candidate model as our final model because it provided more stable performance on unseen data.

- Quadratic Weighted Kappa Scores: 0.766632

The results is slightly lower than 0.854 on the validation set but is still acceptable.

6. SUMMARY AND FUTURE OF WORK

The performance of the model was good overall. It's most effective on identifying No DR, compared to other classes. The model may be biased to Moderate DR, given other classes (except No DR) tend to be identified as Moderate DR. The good news is people with DR can be confirmed with high probability. The bad news is people in worse stage are optimistically diagnosed as moderate, which will delay their treatment.

In the future, we can focus more on the improvement of identification of Severe and Proliferative Classes.

1. We may consider to utilize the techniques of image data augmentation, in purpose of expanding the sizes of these two classes.
2. We can learn from the methods of improving the images quality to reduce the noises or enhance the characteristics of the different categories.

3. Lastly, there are many outstanding algorithms for image classification, such as EfficientNet and DenseNet. It’s always good to explore.