

# **ANALYSIS OF DEEP LEARNING ALGORITHMS FOR DIABETIC RETINOPATHY**

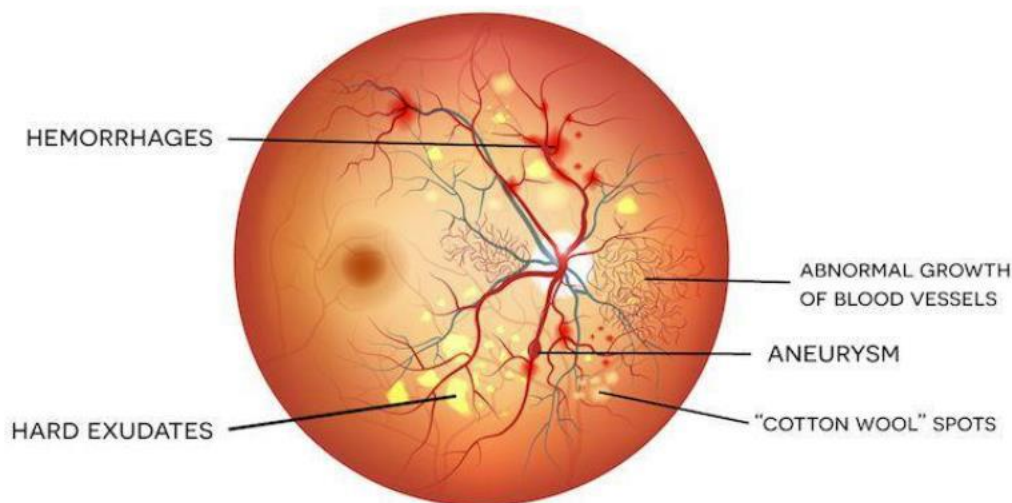
**TEAM NAME: DRISHTI**

**TEAM NUMBER: 5**

**TEAM MEMBERS: Tejashree S, Shedsale Aakash Sunil, Ayushman Raghuvanshi,  
Debraj Banerjee**

## **Introduction:**

- In India, an estimated 60 million people have diabetes. One serious complication of diabetes is diabetic retinopathy (DR), a major cause of avoidable blindness worldwide. Diabetic retinopathy affects approximately 12% to 18% of patients with diabetes in India.
- This condition is caused when high blood sugar levels cause damage to blood vessels in the retina (the light-sensitive area at the back of your eye). These blood vessels can swell and leak. Or they can close, stopping blood from passing through. Sometimes abnormal, new blood vessels grow on the retina. All of these changes can affect a person's vision leading to various degrees of Diabetic Retinopathy.



- Clinically, retinal examination to detect DR is conducted in either of the two ways:
  - i. Direct or indirect ophthalmoscopy or slit-lamp biomicroscopic examination of the retina
  - ii. Retinal (fundus) photography, including any of the following: 30° to a wide field, mono photography or stereo photography, and dilated or undilated photography.

- Globally, grading of DR is practised as per the ICDR scales, which consists of 5 grades from 0 to 4 based on the severity of DR, where
  - 0: No Diabetic Retinopathy
  - 1: Mild Diabetic Retinopathy
  - 2: Moderate Diabetic Retinopathy
  - 3: Non-proliferative Diabetic Retinopathy
  - 4: Proliferative Diabetic Retinopathy
- Telemedicine is potentially a cost-effective solution to access problems, especially those in rural and remote areas. Patients can have retinal images taken at diabetology offices or primary care clinics and the cases can be reviewed by a remote expert.
- Today, a major impediment in implementing tele-screening in tertiary care centres is the lack of trained graders to grade the fundus photography images sent from the remote clinics. Thus, an automated system to assess the severity of DR can help scale screenings.
- Recent work has demonstrated highly accurate deep-learning algorithms for various medical image classification tasks, including retinal imaging. Specifically for DR, multiple groups have shown that deep learning can be leveraged to produce expert-level diagnoses for grading fundus photography images.

## **Methodology and associated results:**

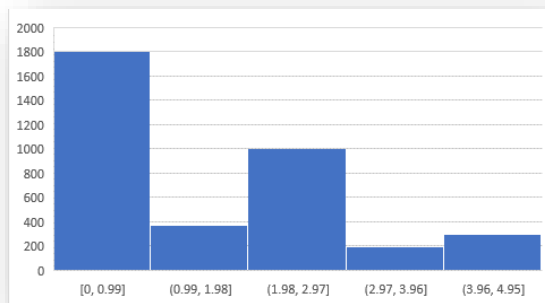
### ***Datasets and DL algorithms used:***

- In this study, we trained few of the existing neural network models for DR classification tasks and we also tried to focus on the explainability aspects of the results we were getting.
- We used the **APTOS-2019 Blindness Detection** Dataset comprising of 3662 labelled fundus images provided by the Aravind Eye Hospital, Madurai and the **IDRiD** dataset comprising of 516 labelled fundus images of the Indian population.
- Some of the neural networks we experimented with are: Resnet-50, Resnet-101, Resnet-18, Resnet-34, Inception-v4, VGG-16 and Alexnet.
- In the further sections, we will try to answer some of the questions we posed and the results that we obtained and the analysis performed on them.

## Q. How does variation in training data affect model performance?

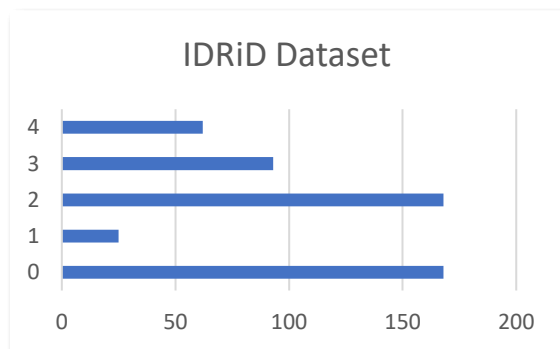
Before we jump into answering this question, let us analyse the datasets at hand.

Variation in data points in each class of APTOS 2019 Blindness Detection Dataset:



Grade	Datapoints
0	1805
1	370
2	999
3	193
4	295

- It can be observed that the distribution of data points in each class is skewed. We are interested in analysing how this affects model performance.
- Now, let us examine the variation in the IDRiD Dataset:



Grade	Datapoints
0	168
1	25
2	168
3	93
4	62

Even here, we can see some skewness in the distribution of the data points. So, we prepared 2 training datasets from APTOS and IDRiD datasets.

- In the first training dataset, we merged data points from both datasets except those in the '0' category. We considered the '0' category images from the APTOS dataset alone. This reduced some skewness in our data distribution.
- In the second training dataset, we randomly chose up to 500 images from the combined datasets of APTOS & IDRiD. This reduced the skewness in data distribution significantly but at the same time, it reduced the total number of data points available for training as well.

**First Training Dataset**

Grade	Datapoints	Normalised values
0	1805	1.0
1	395	0.22
2	1167	0.65
3	286	0.16
4	357	0.20

**Second Training Dataset**

Grade	Datapoints	Normalised values
0	500	1.0
1	395	0.79
2	500	1.0
3	286	0.572
4	357	0.714

- We split our datasets into train and validation sets in the ratio of 85:15 respectively and trained the Alexnet model with Adam optimiser for 300 epochs.
- We curated a test dataset consisting of 100 images from each grade to test the performance of the model.

**Results:**

- Accuracy values for Alexnet models were somewhat low (~40%) but the training times for this model were less as compared to other models. So, we picked this model for epoch-wise performance analysis.
- Epoch-wise analysis of both the models showed that model-1 in spite of having more skewness was able to learn faster and better than model-2.
- Note that in both training datasets the number of images in grades '1', '3', and '4' are the same. Even then, somehow model-1 was able to better predict images belonging to this group than that of model-2. Model-2 mostly learnt to classify images into either grade '0' or '2'.
- This prompted us to use the bigger training dataset-1 in spite of having a higher skewed data distribution for further analysis on bigger models to achieve better metrics.

**Training ResNet-18 and ResNet-34 for DR classification:**

- With insights on training datasets that we obtained from our previous exercise, we now move on to training more complex neural network models for better performance.
- We chose Resnet-18 and Resnet-34 for our training. We applied concepts of transfer learning and trained only the last few layers due to hardware constraints at our end.

- Images are taken in different lighting conditions and this may affect our training. In order to make our model robust to variations in lighting, we applied **Ben-Graham's Lighting** pre-processing technique, which highlighted features of our interest such as exudates, haemorrhages, etc.
- This resulted in achieving an accuracy of **76%** for the Resnet-18 and **75%** for the Resnet-34 models. This is a significant increase in accuracy from the previous Alexnet model.

### *Ben Graham's Image Lighting Preprocessing methodology:*

The idea behind this method is to high boost the image and shift the image intensity by some positive constant. Let  $I$  be the original dark image.

$$\hat{I} = H * I$$

$$I_L = \alpha(I - \hat{I}) + \beta$$

$H$  is a low-pass filter. We pass the original image through this low pass filter to get low pass image. Then we subtract this low spatial frequency, almost uniform image from the original image to get the high-frequency component, scale it appropriately and shift its intensity for brightness. By this, we have successfully added weights to those parts that are finer, less uniform and are better indicators for grading. Here,  $\alpha$  and  $\beta$  are hyper-parameters.

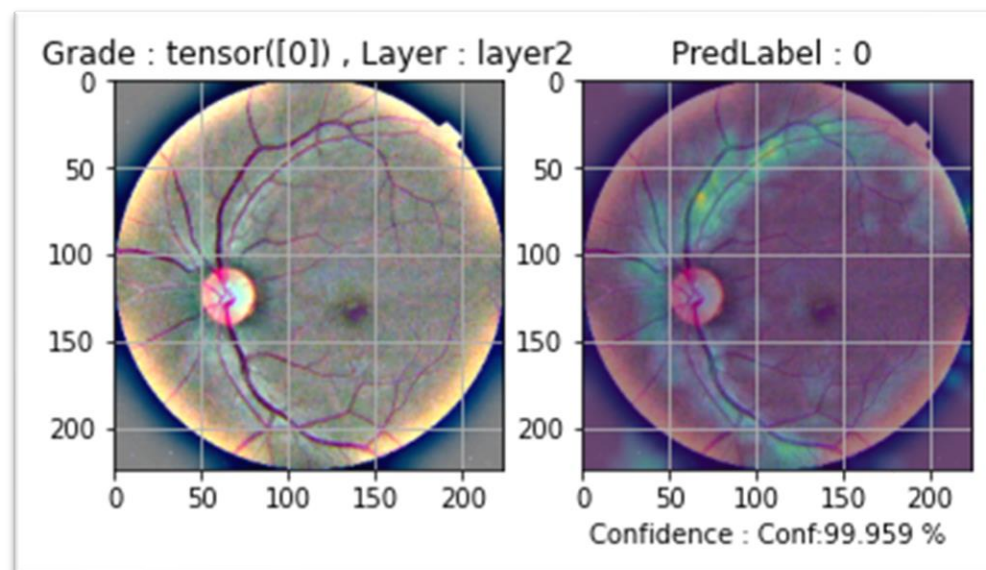
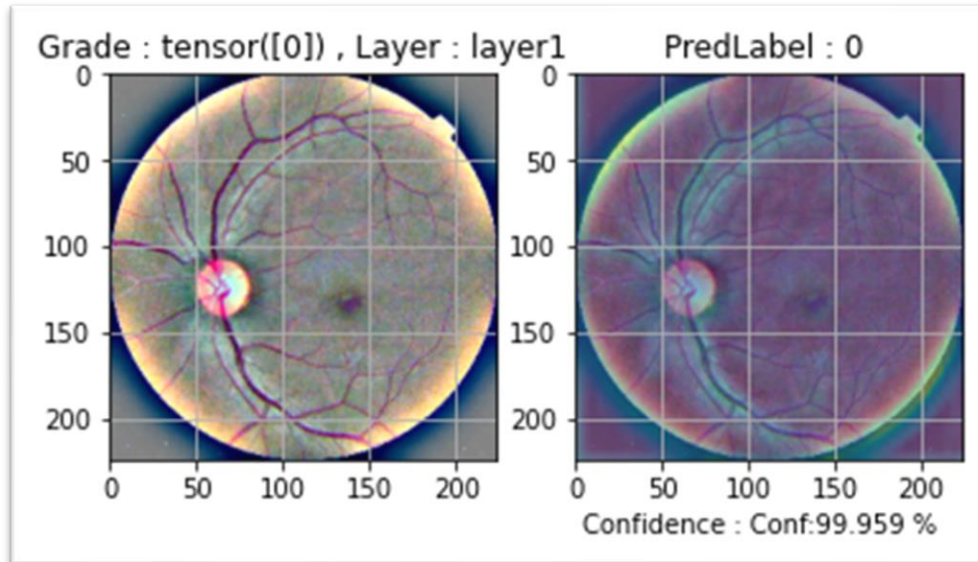
#### Images before and after Ben Graham's pre-processing



- To get insights into how the model is learning, we made use of Grad-CAM outputs at the end of each block of layers.
- The Grad-CAM technique utilises the gradients of the classification score with respect to the final convolutional feature map, to identify the parts of an input image that most impact the classification score. The places where this gradient is large are precisely the places where the final score depends most on the data.
- In order to get an overall view of the learnings of the model, we take an average over the 4 Grad-CAM outputs, which gives information about the average learning and helps in visualising the overall learning of the model.
- Grad-CAM outputs for each block of layers and the averaged Grad-CAM outputs are visualised in the following section.

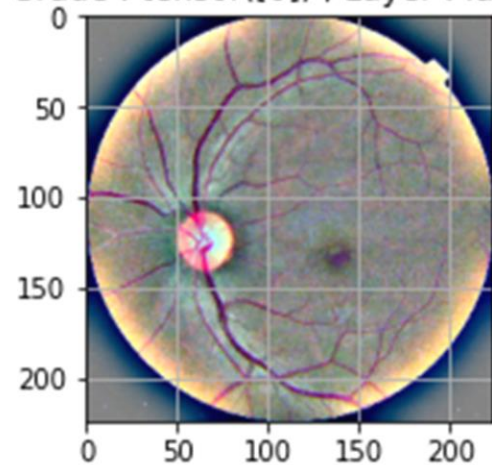
- **Results:** The Resnet-18 & Resnet-34 models are able to predict pre-processed images with higher confidence levels and are also able to detect the areas of lesions, haemorrhages etc.

Grad-CAM outputs at the end of each block of layers of a grade '0' image:

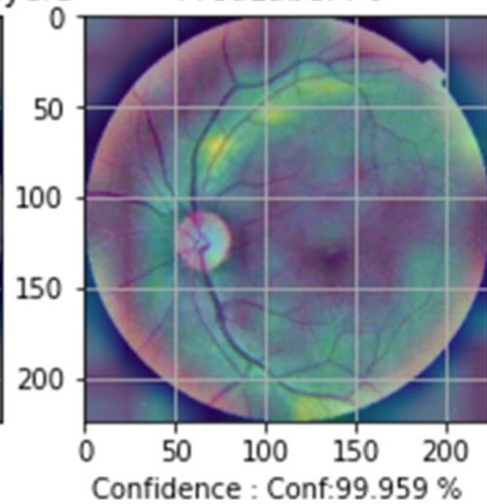




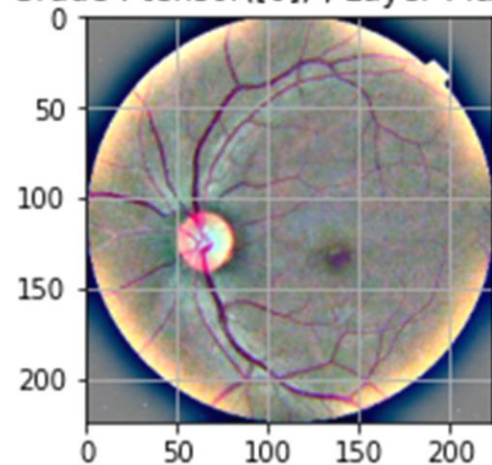
Grade : tensor([0]) , Layer : layer3



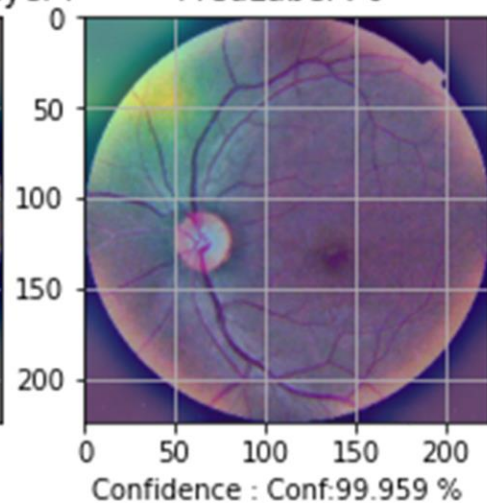
PredLabel : 0



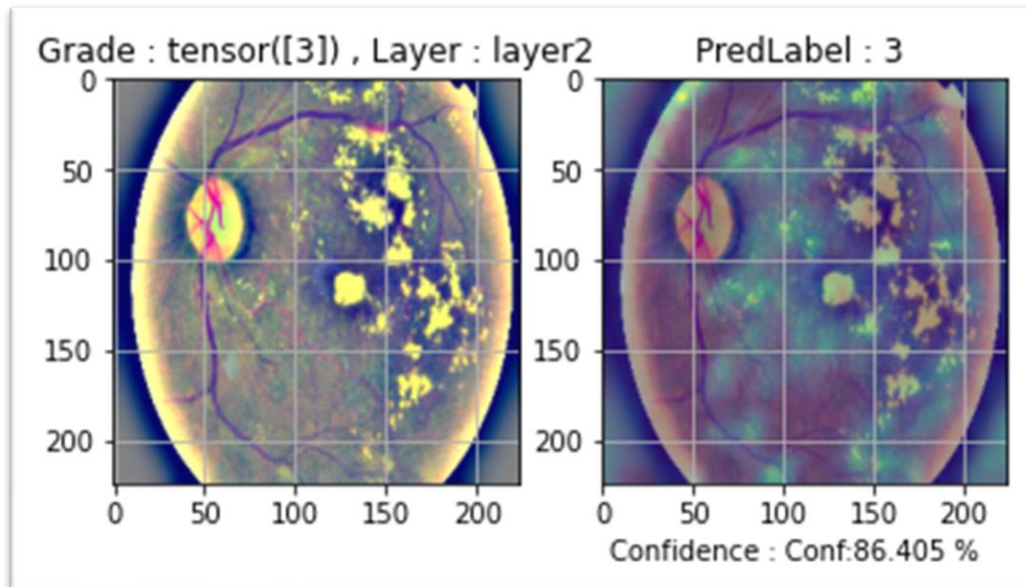
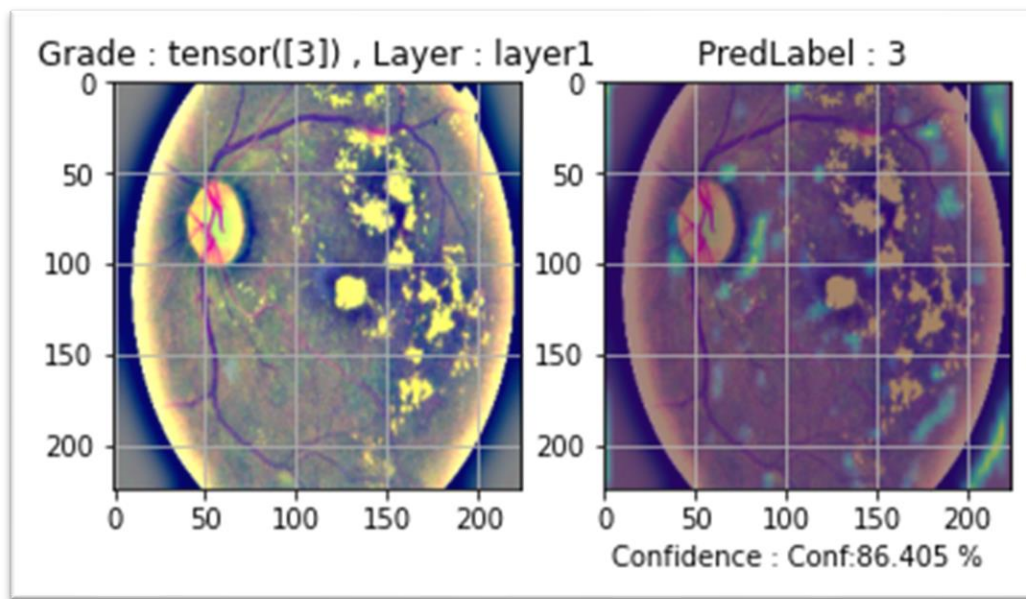
Grade : tensor([0]) , Layer : layer4



PredLabel : 0

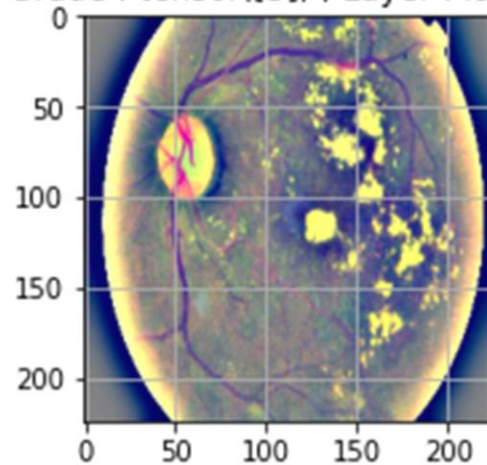


Grad-CAM outputs at the end of each block of layers of a grade '3' image:

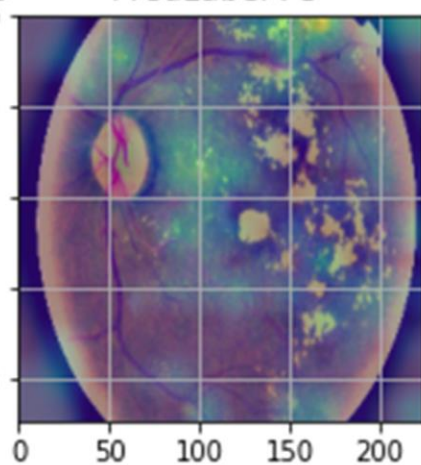




Grade : tensor([3]) , Layer : layer3

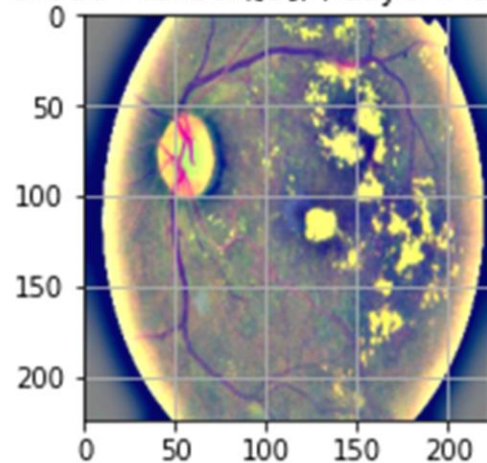


PredLabel : 3

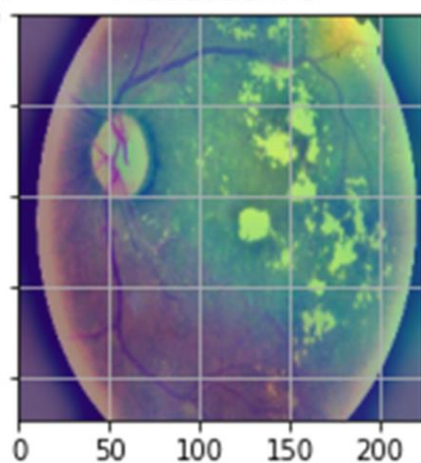


Confidence : Conf:86.405 %

Grade : tensor([3]) , Layer : layer4

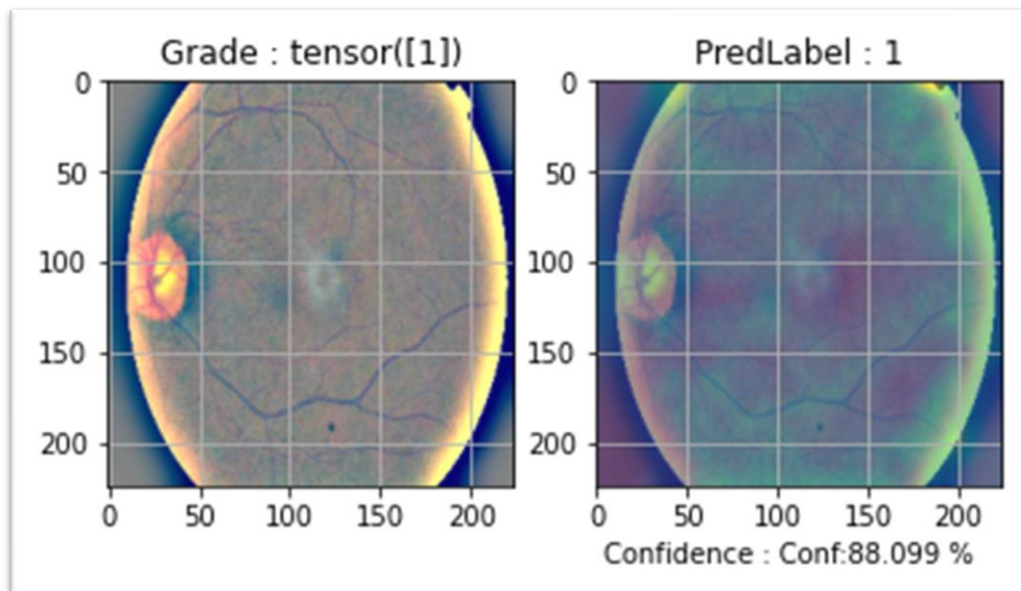
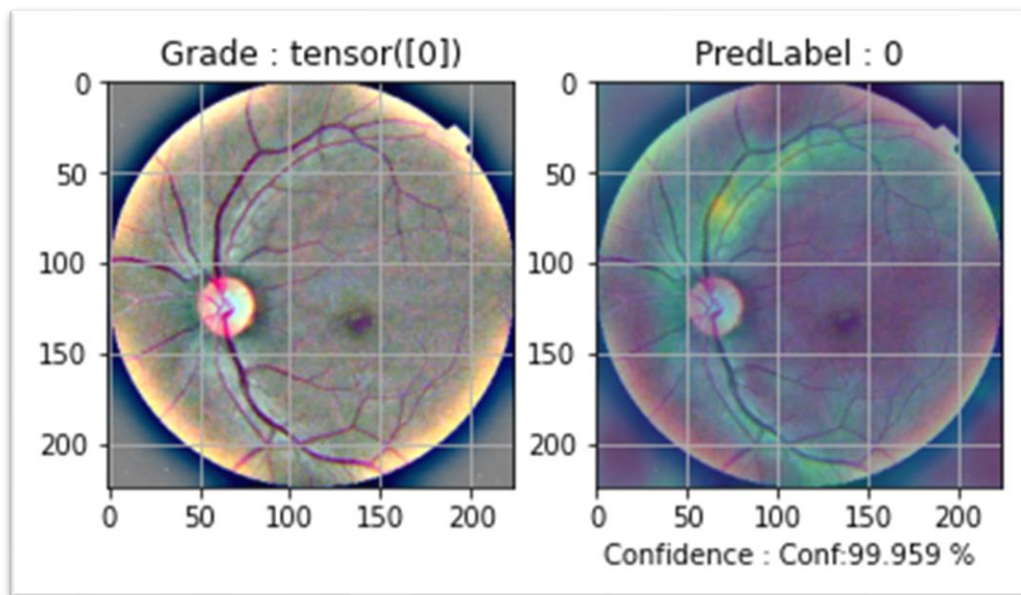


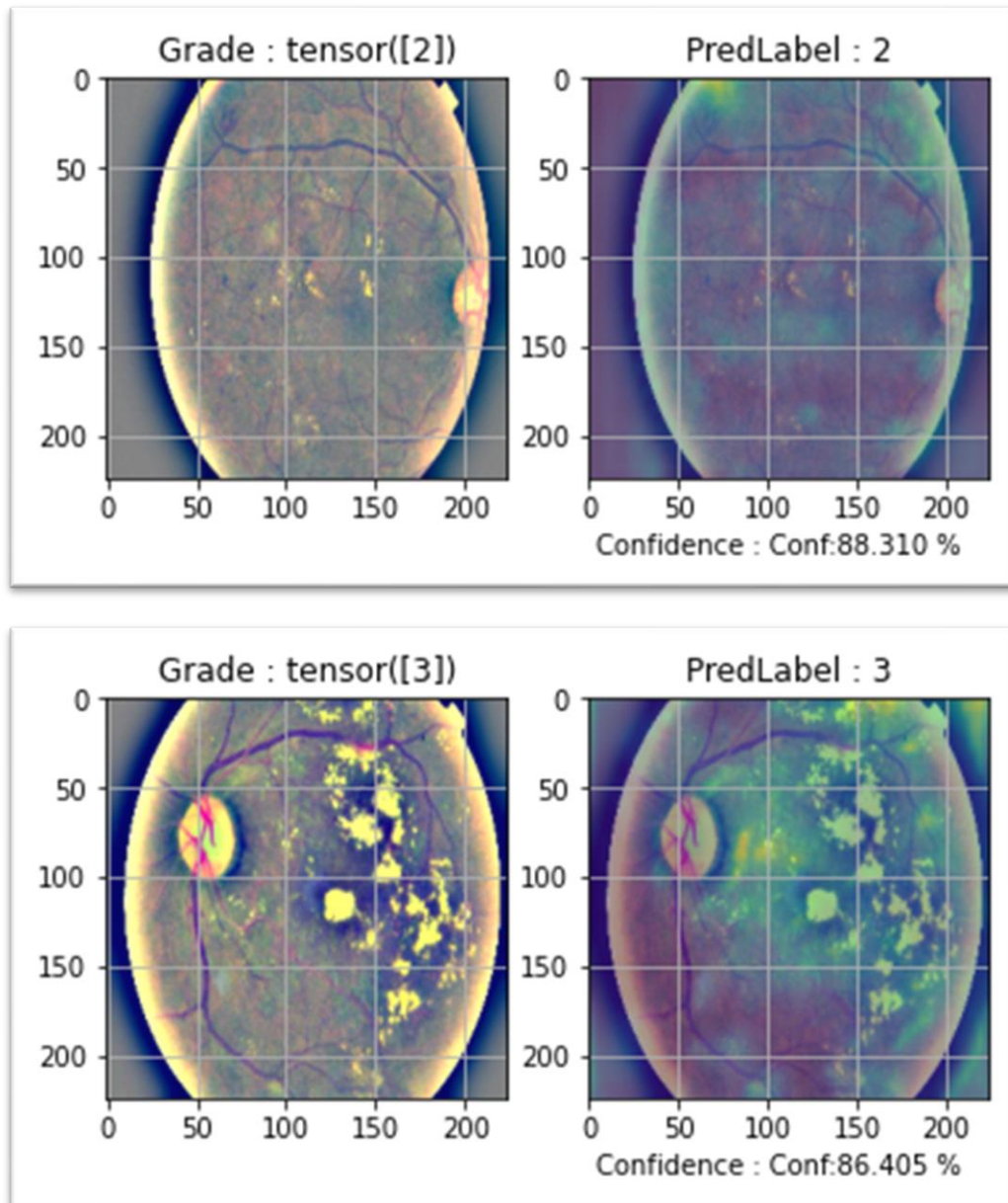
PredLabel : 3



Confidence : Conf:86.405 %

Averaged Grad-CAM Outputs for images belonging to different classes:





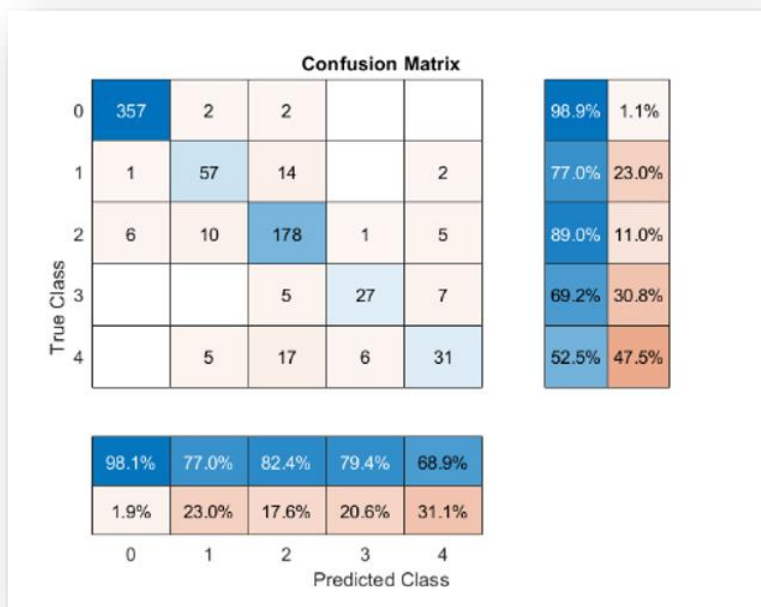
### Training ResNet-50 and ResNet-101 for DR Classification:

- Next, we experimented with deeper models with even higher layers. We chose ResNet-50 and ResNet-101 for this.
- We used the APTOS dataset without any pre-processing techniques applied to the images. We used an 80:20 train:test split on the APTOS dataset for this exercise.
- We got better accuracies than the previous models. We achieved **88.40%** accuracy for the ResNet-50 model and **88.68%** accuracy for the ResNet-101 model.
- When we observed the Grad-CAM outputs of every layer, we noticed that the initial few layers were learning the visual features of our interest. So, to investigate this, we took

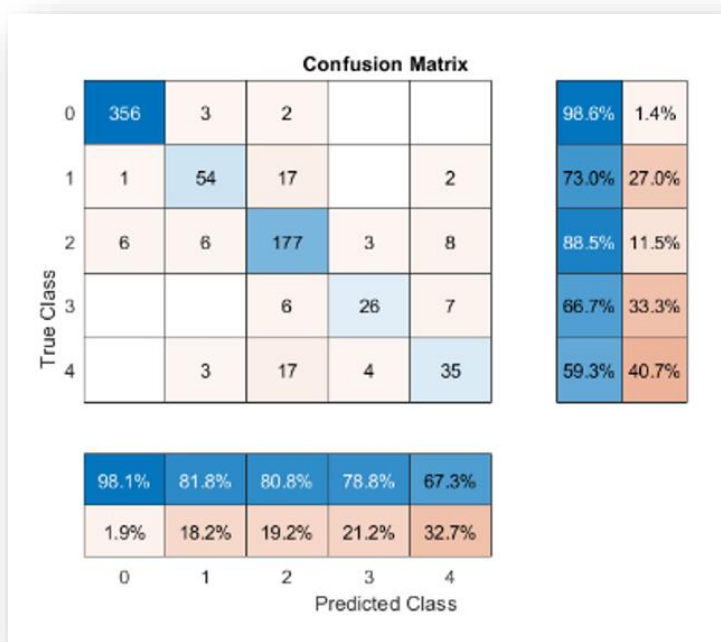
the outputs of the 15th layer of ResNet-101 and the 16th layer of ResNet-50 and connected them to a fully connected layer and took predictions and tested them for accuracy.

## Confusion Matrices:

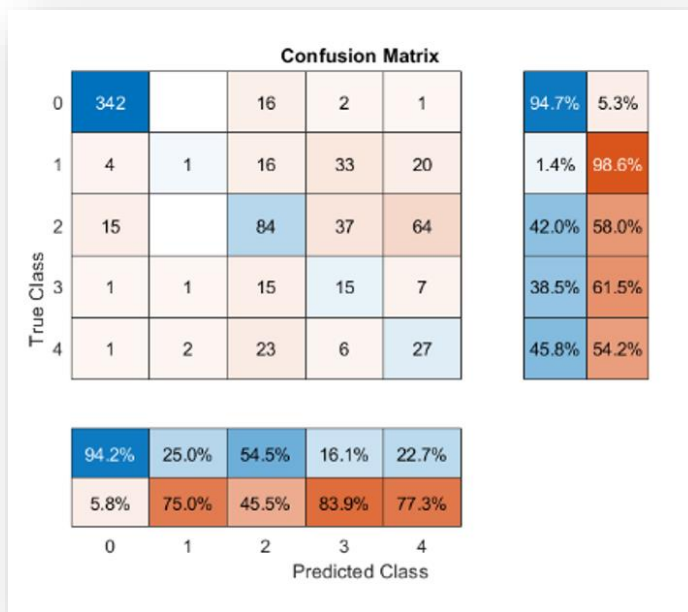
- ResNet-101 all layers accuracy **88.68%**



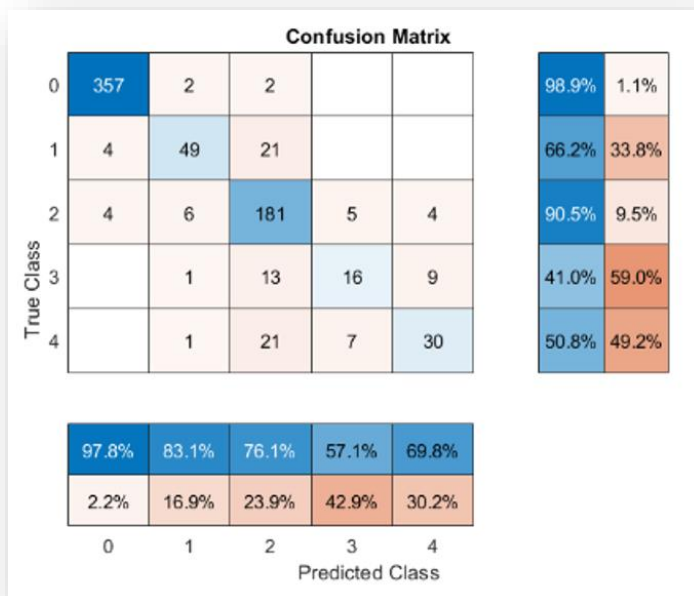
- ResNet-50 all layers accuracy **88.40%**



- ResNet-50 first 16 layers accuracy **63.98%**



- ResNet-101 first 15 layers accuracy **86.36%**



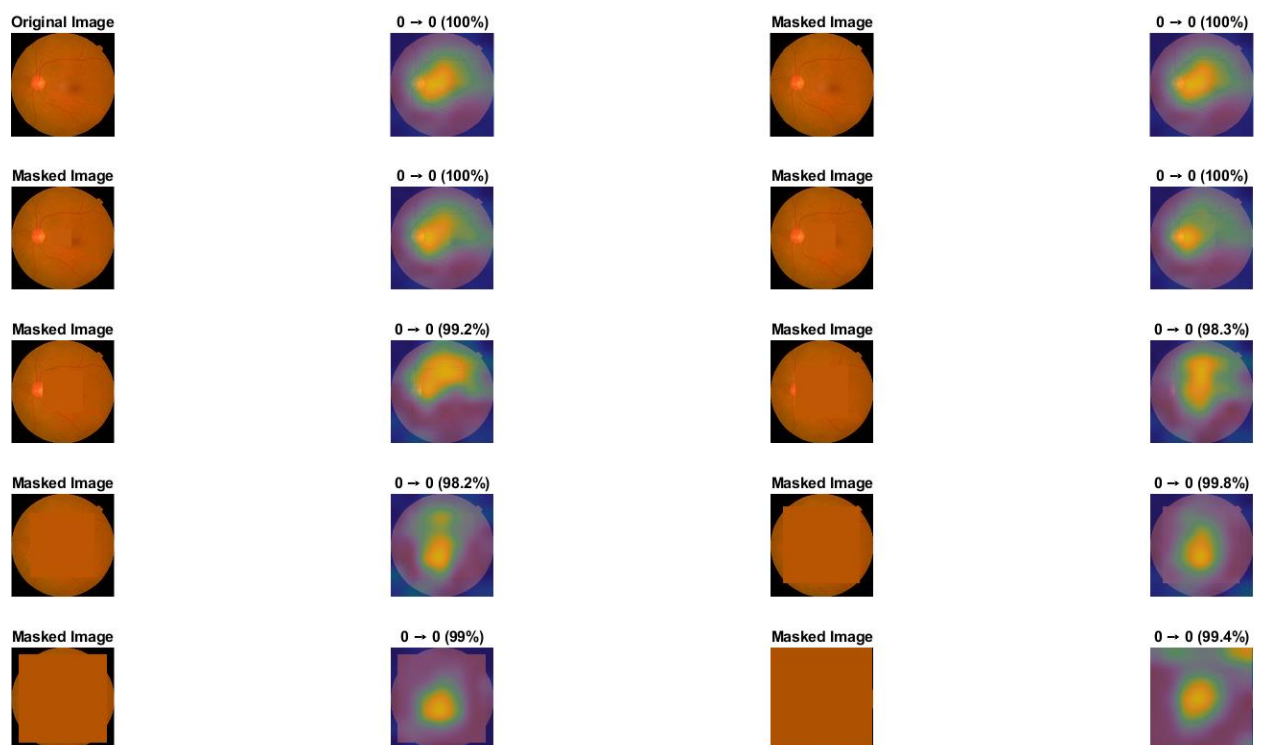
- As observed in the Grad-CAM visualisations, the above accuracy results prove that the first few layers were learning most of the important features for DR Classification successfully.



## Q. Which areas of the fundus image are most affecting the prediction?

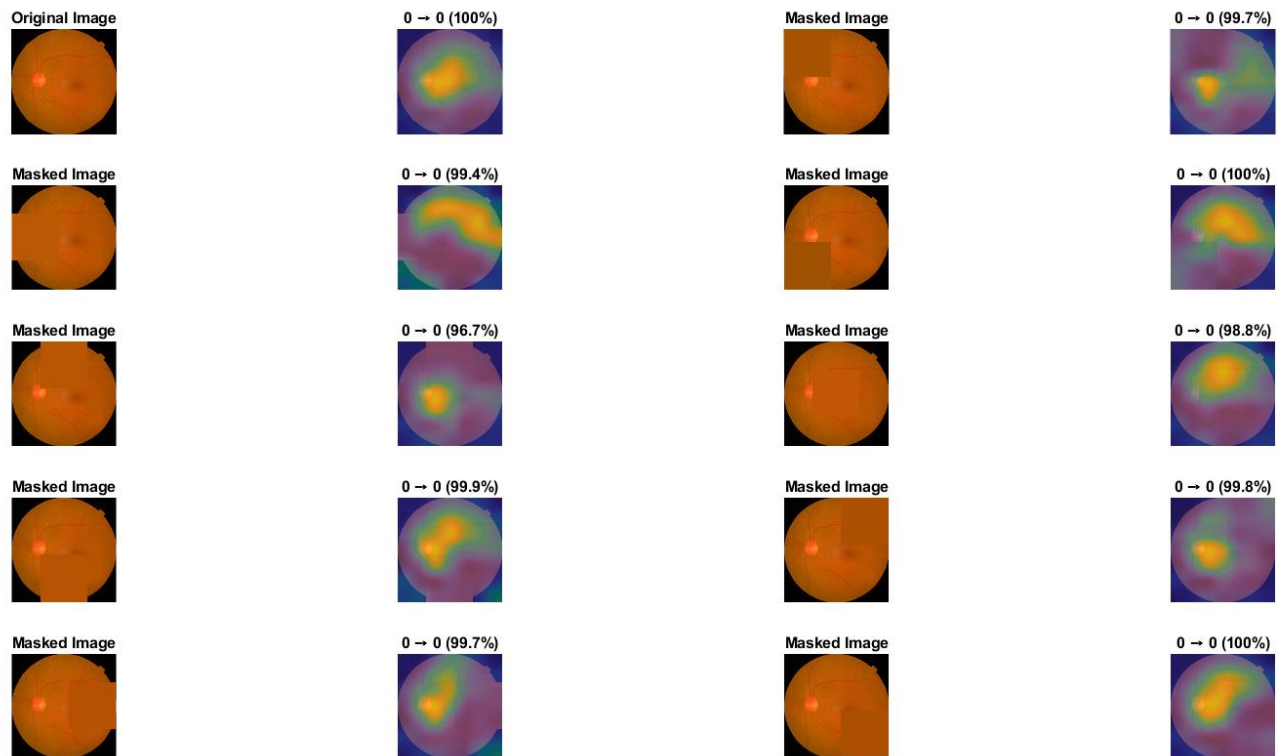
- Now, that we received good accuracy using ResNet-50 and ResNet-101 models, let us dive into discovering which areas within the fundus image most affect the predictions. To answer this, we took test images whose grades were predicted by the model accurately.
- Then, we used varying patch sizes ranging from 10 x 10 to 100 x 100 as a mask which replaced actual pixel values with the median value of the pixels under the patch. By shifting this patch mask throughout the image, we generated a set of test images which were fed to our model for prediction.
- We carried out this exercise to find those areas whose masking resulted in a dip in the prediction accuracy levels. These areas would be of interest as the features present in these areas would likely be the ones that the model had learnt for classification.

ResNet-50 all layers for **varying patch size** for a fixed class of image (grade '0'):

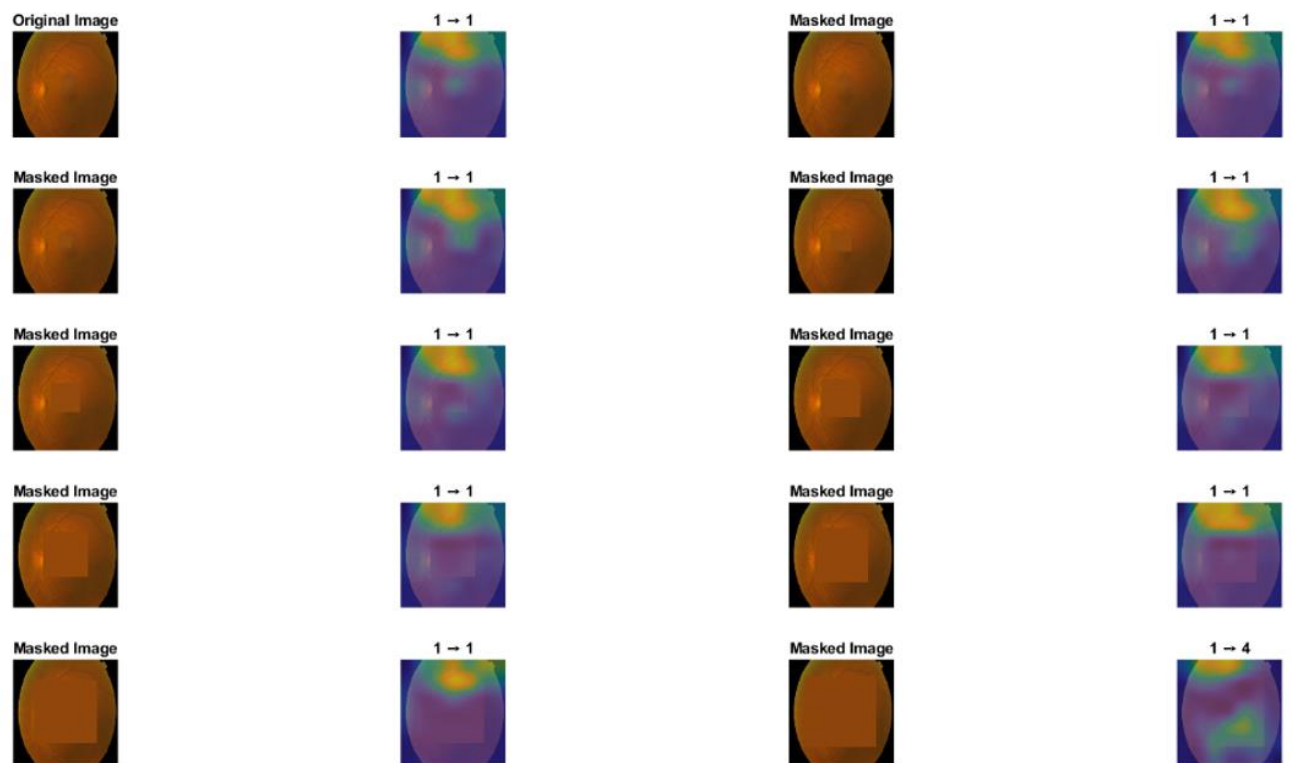




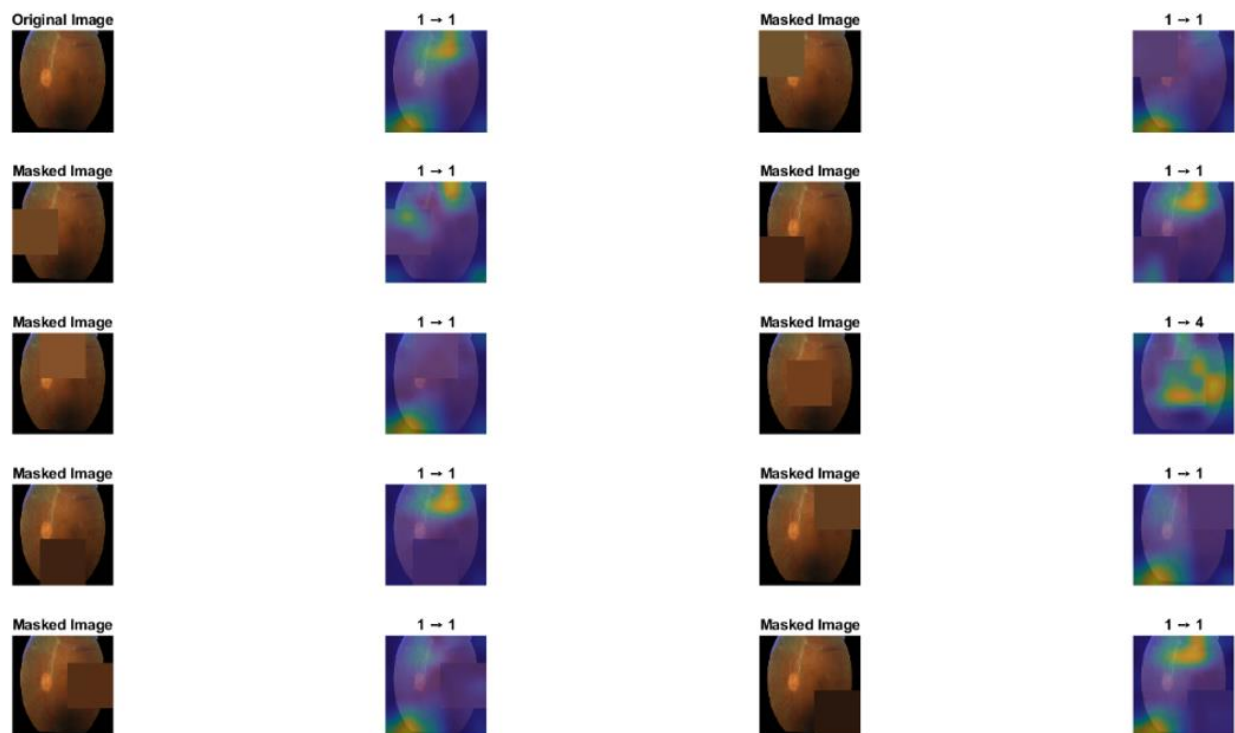
ResNet-50 all layers for **different patch locations** for a fixed class of image (grade '0'):



ResNet-50 all layers for **varying patch size** for a fixed class of image (grade '1'):

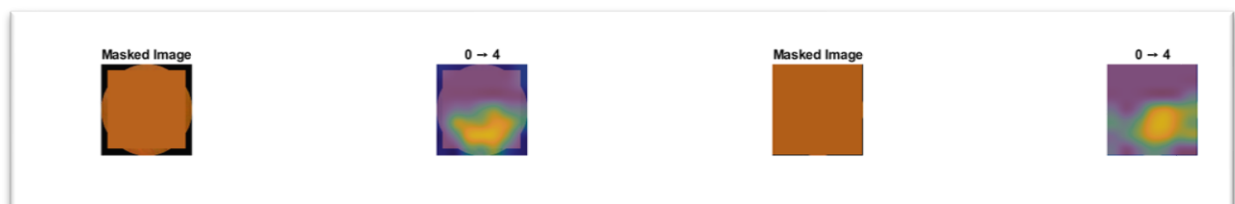


ResNet-50 all layers for **different patch locations** for a fixed image (grade '1'):



### **Observations:**

- As the input data contains a disproportionate amount of grade '0' data (about 49% of total data), the model kind of overfits the data as a result, we get almost 100% accurate predictions even if the patch size is increased further. Only for some extreme cases as shown below, the model predicts incorrect class.
- For grade '1' images, we observe that as the patch size increased or when the location of the patch changed, the model classified wrongly or the confidence of the prediction reduced, hinting that the latent features are more localised for other grades than that of grade '0' images.
- Also, we observed that when the patch masked areas around the middle of the fundus image comprising of the fovea, optic disc and surrounding areas, the model predicted wrongly, leading us to further investigate which of these areas contributed the most features for the classification.



Q. Which are the important areas: Fovea, Optic Disc, centre or all?

- From the insights we got from the previous exercise, let us figure out whether the fovea, optic disc or some area in between them or all of them contribute to the classification task in this section.
- In order to achieve this, we cropped our images centred at the fovea, optic disc and the area in between them and trained ResNet-34 and ResNet-50 models on these cropped images.
- Then we presented the corresponding cropped areas of the test images to see how the models performed. The results obtained are shown below.

Image showing the marking of fovea (F) and optic disc (OD):

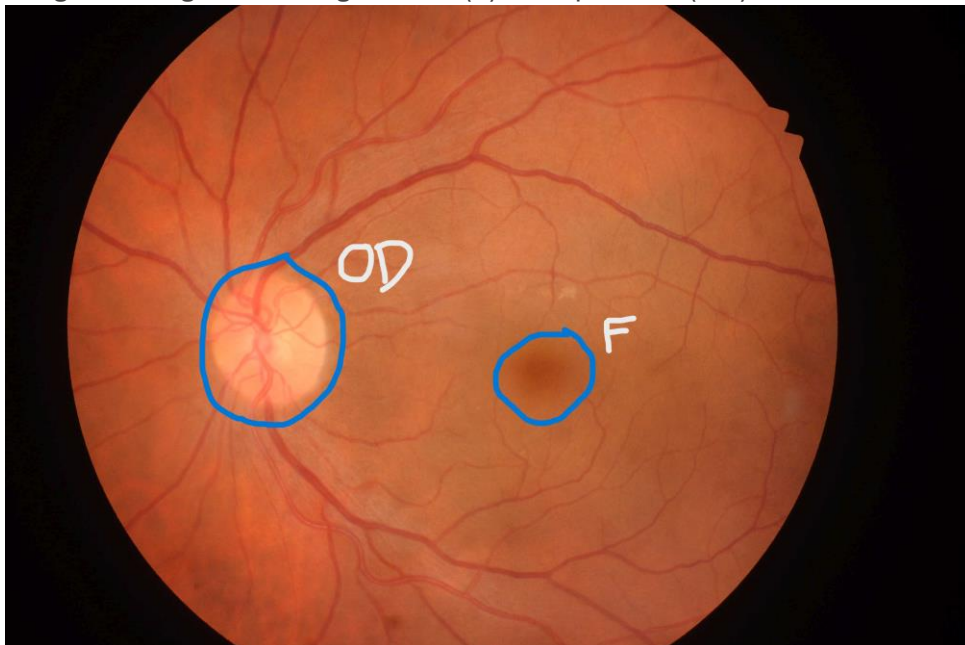


Image cropped around fovea:



Image cropped around the optic disc:



Image cropped around the centre of the fundus image:



Performance metrics of model trained on cropped fovea images:

Fovea	Accuracy	Recall	F1 score
ResNet-50	0.84	0.85	0.85
ResNet-34	0.82	0.87	0.81

Performance metrics of model trained on cropped optic disc images:

Optic Disc	Accuracy	Recall	F1 score
ResNet-50	0.77	0.77	0.77
ResNet-34	0.79	0.79	0.79

Performance metrics of model trained on cropped fovea images:

Center eye area	Accuracy	Recall	F1 score
ResNet-50	0.79	0.78	0.77
ResNet-34	0.83	0.83	0.83

By the above measurements, we can conclude that fovea, optic disc and eye center were all important areas that contributed to the classification, with croppings centred around fovea providing higher performance metrics, closer to the performance associated with full fundus images.

## **Conclusion:**

- Through this project, we observed that a larger dataset with higher skewness in data distribution trained the model better and faster than a smaller dataset with lesser skewness in data distribution.
- Ben-Graham Lighting pre-processing technique resulted in better model performance. The model was also able to learn the features of interest such as haemorrhages, hard exudates etc.
- ResNet-50 and ResNet-101 models gave the best performance out of all the models we experimented with.
- The first few initial layers of both ResNet-50 and ResNet-101 yielded high accuracy of predictions, especially that of ResNet-101 which was almost equal to the final accuracy obtained at the end of all layers.
- The model was robust to almost all types of masking for grade '0' images owing to the higher number of data points in the training dataset, whereas it was somewhat sensitive to the maskings near the middle portion of the fundus image.
- Images centred around the fovea, optic disc and centre contributed significantly to the classification, among which those centred around the fovea produced very close performance similar to that of full fundus images.



## **Bibliography**

1. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216
2. Gulshan V, Rajan RP, Widner K, et al. Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. *JAMA Ophthalmol*. 2019;137(9):987–993. doi:10.1001/jamaophthalmol.2019.2004
3. M. Karakaya and R. S. Aygun, "Retinal Biomarkers for Detecting Diabetic Retinopathy Using Smartphone-Based Deep Learning Frameworks," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095574.
4. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, Meriaudeau F. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Data*. 2018; 3(3):25. <https://doi.org/10.3390/data3030025>
5. <https://www.kaggle.com/c/aptos2019-blindness-detection>