Diabetic Retinopathy Detection using Deep Learning and Image processing

Abstract

Diabetic retinopathy is a leading cause of blindness among working-age adults. Early detection of this condition is critical for a good prognosis. In this paper, we demonstrate the use of convolutional neural networks (CNNs) on color fundus images for the recognition task of diabetic retinopathy and it's grading. Our network models through a two-level ensemble approach achieved new benchmark results on the validation set with our test metric Quadratic Weighted Kappa(QWK) crossing the best public available QWK score achieving a value of 0.9417. We additionally explored multinomial classification models, and demonstrate that errors primarily occur in the misclassification of mild disease as normal due to the CNNs inability to detect subtle disease features. We discovered that preprocessing with contrast limited adaptive histogram equalization and ensuring dataset fidelity by expert verification of class labels improves recognition of subtle features, furthermore, we investigate other image preprocessing techniques such as median subtraction, and Ben Graham's technique, with best results obtained using Ben Grahams's preprocessing. Transfer learning using pre-trained Deep CNN architectures Resnet and Efficient-Net models trained on ImageNet improved peak test set QWK score to 0.9417 on 5 class classification models using an ensemble approach at two levels, respectively.

Introduction:

Among individuals with diabetes, the prevalence of diabetic retinopathy is approximately 28.5%(14) in the United States and 18%(16) in India. Most guidelines recommend annual screening for those with no retinopathy or mild diabetic retinopathy, repeat examination in 6 months for moderate diabetic retinopathy, and an ophthalmologist referral for treatment evaluation within a few weeks to months for severe or worse diabetic retinopathy or the presence of referable diabetic macular edema, known as clinically significant macular edema. Referable diabetic retinopathy has been defined as moderate or worse diabetic retinopathy or referable diabetic macular edema, given that recommended management changes from yearly screening to closer follow-up at moderate disease severity. Retinal photography with manual interpretation is a widely accepted screening tool for diabetic
retinopathy, with a performance that can exceed that of in-person dilated eye examinations. Automated grading of diabetic retinopathy has potential benefits such as increasing efficiency, reproducibility, and coverage of screening programs; reducing barriers to access; and improving patient outcomes by providing early detection and treatment. To maximize the

clinical utility of automated grading, an algorithm to detect referable diabetic retinopathy is needed. Machine learning (a discipline within computer science that focuses on teaching machines to detect patterns in data) has been leveraged for a variety of classification tasks including automated classification of diabetic retinopathy.
However, much of the work has focused on "feature engineering," which involves computing explicit features specified by experts, resulting in algorithms designed to
detect specific lesions or predicting the presence of any level of diabetic retinopathy. Deep learning is a machine learning technique that avoids such engineering by learning the most predictive features directly from the images given a large
data set of labeled examples. This technique uses an optimization algorithm called back-propagation to indicate how a machine should change its internal parameters to best predict the desired output of an image.

Need for such a system:

Approximately four hundred and twenty million people worldwide have been diagnosed with diabetes mellitus. The prevalence of this disease has doubled in the past 30 years 24 and is only expected to increase, particularly in Asia7. Of those with diabetes, approximately one-third are expected to be diagnosed with diabetic retinopathy (DR), a chronic eye disease that can progress to irreversible vision loss 8. Early detection, which is critical for a good prognosis, relies on skilled readers and is both labor and time-intensive. This poses a challenge in areas that traditionally lack access to skilled clinical facilities. Moreover, the manual nature of DR screening methods promotes widespread inconsistency among readers. Finally, given an increase in the prevalence of both diabetes and associated retinal complications throughout the world, manual methods of diagnosis may be unable to keep pace with demand for screening services12.

Automated techniques for diabetic retinopathy diagnoses are essential to solving these problems. While deep learning for binary classification, in general, has achieved high validation accuracies, multi-stage classification results are less impressive, particularly for early-stage disease.

In this paper, we introduce an automatic robust DR grading system capable of classifying images based on disease pathologies from five severity levels. A convolutional neural network (CNN) convolves an input image with a defined weight matrix to extract specific image features without losing spatial arrangement information.

Related Work and existing systems:
Diagnosis of pathological findings in fundoscopy, a medical technique to visualize the retina, depends on a complex range of features and localization within the image. The diagnosis is

particularly difficult for patients with early-stage diabetic retinopathy as this relies on discerning the presence of microaneurysms, small saccular outpouching of capillaries, retinal hemorrhages, ruptured blood vessels—among other features—on the fundoscopic images.

Computer-aided diagnosis of diabetic retinopathy has been explored in the past to reduce the burden on ophthalmologists and mitigate diagnostic inconsistencies between manual readers16. Automated methods to detect microaneurysms and reliably grade fundoscopic images of diabetic retinopathy patients have been active areas of research in computer vision19. The first artificial neural networks explored the ability to classify patches of the normal retina without blood vessels, normal retinas with blood vessels, pathological retinas with exudates, and pathologic retinas with microaneurysms. The accuracy of being able to detect microaneurysms compared to normal patches of the retina was reported at 74% 10.

Past studies using various high bias, low variance digital image processing techniques have performed well at identifying one specific feature used in the detection of subtle diseases such as the use of a top-hat algorithm for microaneurysm detection 17,23,16. However, a variety of other features besides microaneurysms are efficacious for disease detection.

Additional methods of detecting microaneurysms and grading DR involving k-NN5,20, support vector machines 22, and ensemble-based methods 6 have yielded sensitivities and specificities within the 90% range using various feature extraction techniques and preprocessing algorithms.

Previous CNN studies14,11 for DR fundus images achieved sensitivities and specificities in the range of 90% for binary classification categories of normal or mild vs moderate or severe on much larger private datasets of 80,000 to 120,000 images. However, accuracy measures for the detection of four classes of DR, that is no DR (R0), mild (R1), moderate (R2), and proliferative (R3), and severe(R4) depend nontrivially on disease graded class collection ratios. While R0 and R4 stages are capable of achieving high sensitivity, but R1, R2, and R3 computed recall rates are often low. Experiments from publicly available datasets suggest this is primarily attributable to the relative difficulty of detecting early-stage DR. Furthermore, current accuracies for R1, R3 and R2 stages are reported at 0% and 41%, respectively.

Currently, the error rate in the detection of an eye threatening retinopathy is 40% for optometrists and 35% for an ophthalmologist as reported by NCBI(National center for biotechnology information).

The latest automated systems based on deep learning that is available for DR detection although almost all of them have specificity in the range of(90%-97%) and sensitivity in the range of(91%-96%) are as follows:

Valentina balemmo's(May 2019)[3]- AI-based image classifier for DR detection is the latest deep learning-based model with an ensemble approach.

In another work by Wei Zhang, Jie Zhong(March 2019)[4] and others develop another ensemble-based image classifier for DR detection that has produced very interesting results.

Both of these works are the most recent and the best results obtained so far in DR detection, but both of these studies are different because Balemmo uses 3 class labels while Wei Zhang and others use $ class labels for Dr detection and classification.

Dataset used by Balemmo consists of around 76000 images(from 13000 patients) in training and 4500 images for validation which although is quite big and the model thus produced is Robust but it only classifies it into 3 classes.

On the other hand, Wei Zhang and others use a dataset consisting of around 13500 images which are then divided into training, validation and test set, another important point is they use 4 classes for there DR detection system.

Other works in the area of DR detection using deep learning and image processing include [7][8][9][10].

The aim of this study is to produce a Robust model for DR detection over the 5 classes to produce a model that is robust as well as is able to generalize well when faced with different kinds of images taken in different conditions with different kinds of devices.

A Brief Literature Review:

S.NO	Authors	Year	Technique	Quadratic weighted kappa	AUC-ROC Metric	Dataset Used
1	Varun Gulshan, Ph.D.; Lily Peng, MD, Ph.D.; Marc Coram, Ph.D.; Martin C. Stumpe, Ph.D.; Derek Wu, BS; A	2016	Inception V-3.	0.8851(5)	0.991(3 vs All), 0.993(Avg-AUC across all datasets)	Eyepacs-1(15,000, Messidor-2(2000), Google's own DB-1,20,000, APTOS(13000)
2	Daniel Shu Wei Ting, MD, PhD[1,2]; Carol Yim-Lui Cheung, PhD[1,3]; Gilbert Lim, PhD[4]; et al	2017	Inception V-3 and V-4 + CLAHE	0.856(5)	0.971(4 vs All), Avg AUC-0.951	Eyepacs-2(76000),Messidor-2(200)
3	Bellemo, Valentina, et al.	2019	VGG-net	0.831(3)	0.973(2 vs All) and	Eyepacs-2, and Zambian eye

						society(18000)
4	Wei Zhang, Jie Zhong, Shijun Yang, Zhentao Gao	2019	Two-Part ensemble. 1-binary classification then, 2- grading System. 1-Xception,Resnet -50,Inception resnet-v2.	0.8771(4)	0.981(Classifi cation), 0.994(Avg-AUC)	Sichuan Medical Center(15000)

Dataset Description:

Diabetic retinopathy is usually diagnosed by an ophthalmologist and graded into 5 categories as shown:

- 0-No DR.
- 1-Mild DR.
- 2-Moderate DR.
- 3-Severe DR.
- 4-Proliferative DR.
- This grading scheme was designed to tackle the earlier complex procedure for scaling DR diagnosis which was also known as ETDRS(Early Treatment of Diabetic retinopathy Study).

Due to the complexity of ETDRS a new scale was adopted which was easy to understand and implement and also improved diagnosis, this came to be known as the INTERNATIONAL CLINICAL DISEASE SEVERITY SCALE FOR DR [1][2][6].

For this Project we selected Open- Sourced datasets that have been made publicly available:

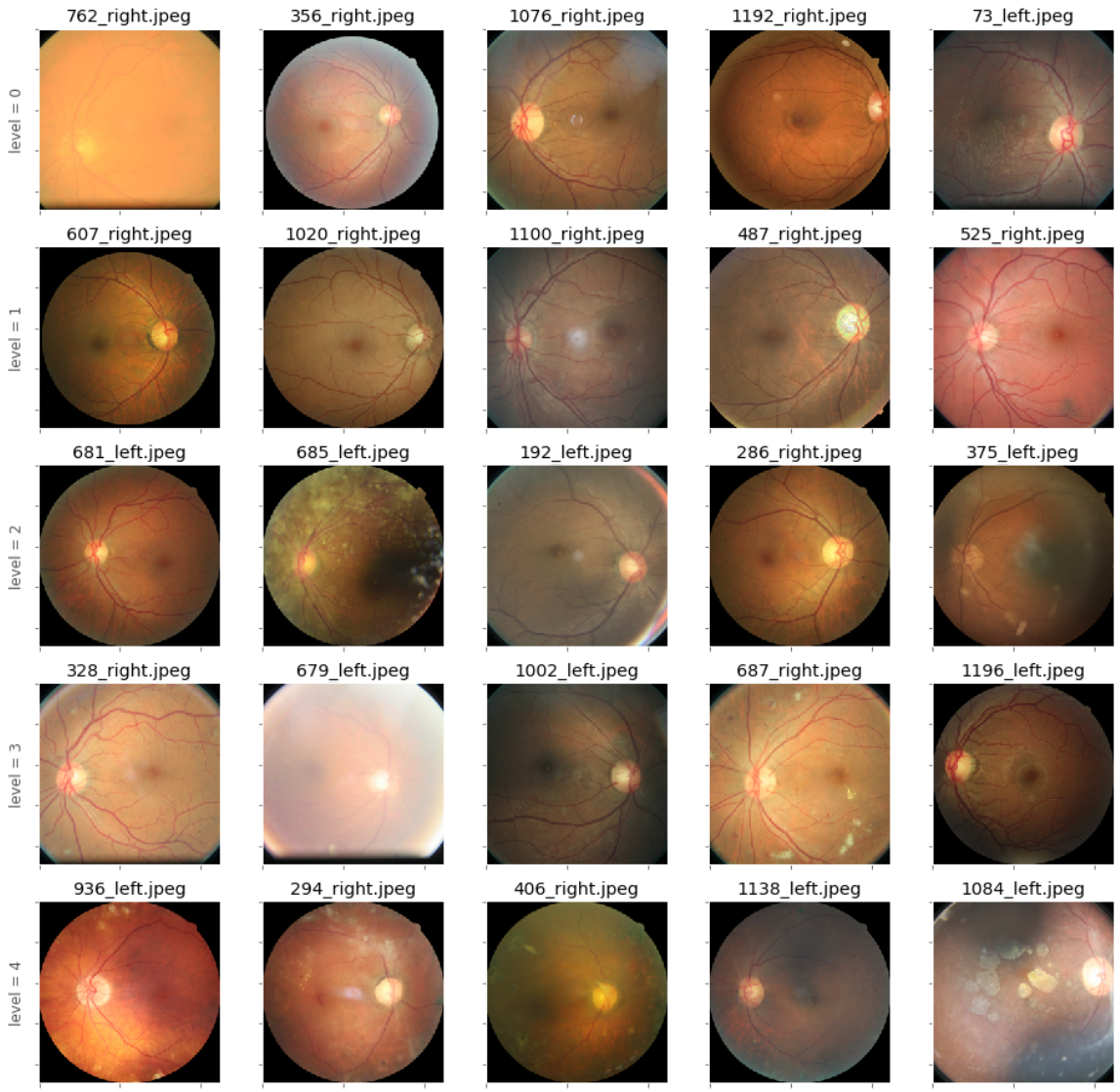
- The First dataset we took from Eyepacs the Eyepacs - 2 dataset containing 1,20,000 images.
- Eyepacs - 1 dataset from Kaggle as part of the DR detection competition 5 years ago. It Contains 84,000 images of 41,254 patients.
- APTOS DR detection Competition, a Kaggle dataset now made publicly available by Aravind Eyecare, It consists of 8,000 images.
- Messidor-2 dataset, It consists of 2000 images.
- All the images used are Retino-fundal images with a fixed resolution of 1024 x 1024 for all the images i.e in total we have 2,14,000 retino-fundal images of the retina which have already been graded by ophthalmologists into 5 categories of DR which are as follows:

- 0-No DR.
- 1-Mild DR.
- 2-Moderate DR.
- 3-Severe DR.
- 4-Proliferative DR.

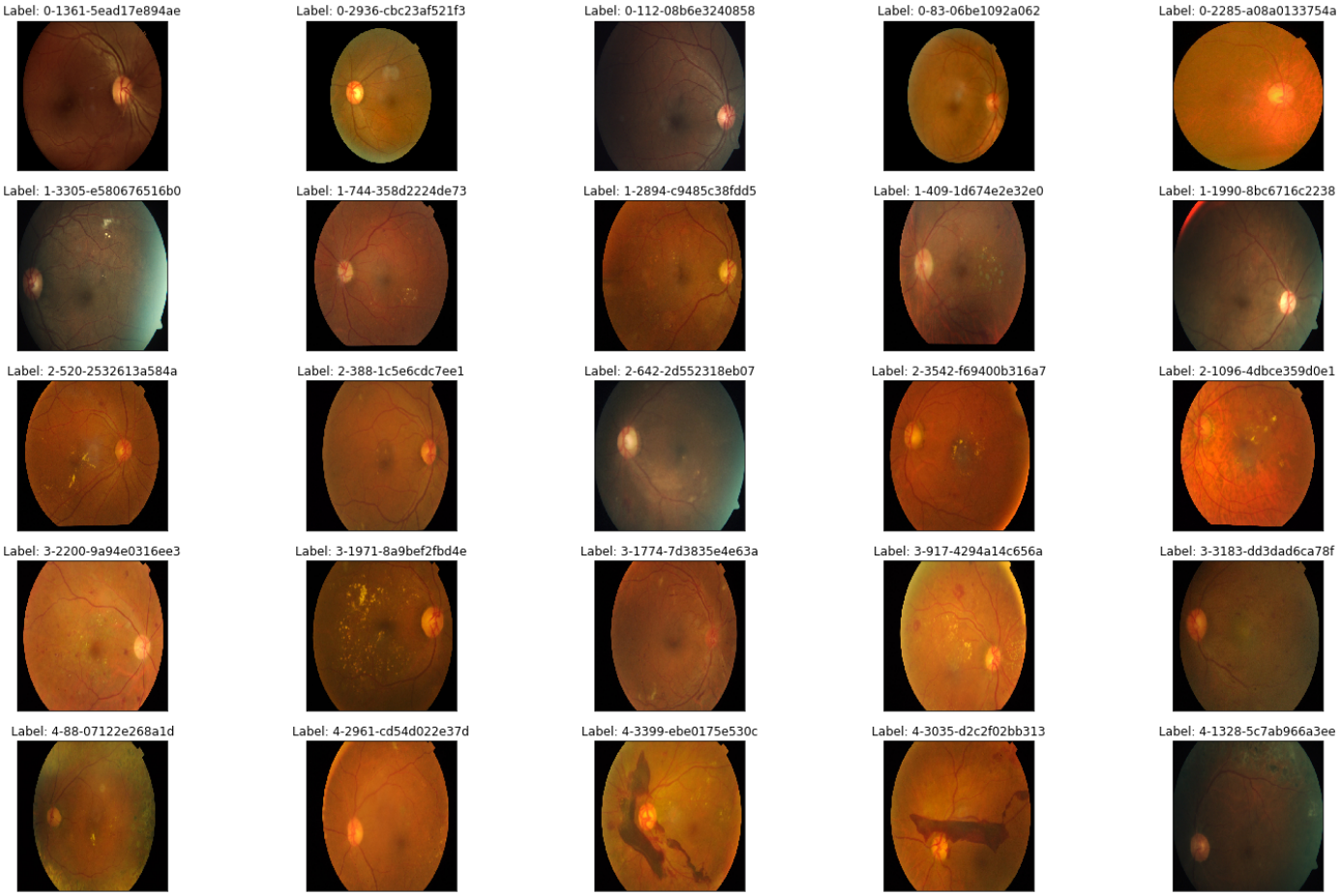
We trained our models on 1,50,000 images then further fine-tuned the models using 40000 images. We validated our models on the remaining 15000 images. We finally tested our models on the remaining 13,000 images using Test Time Augmentation.

Some images from the dataset:

Eyepacs-1:

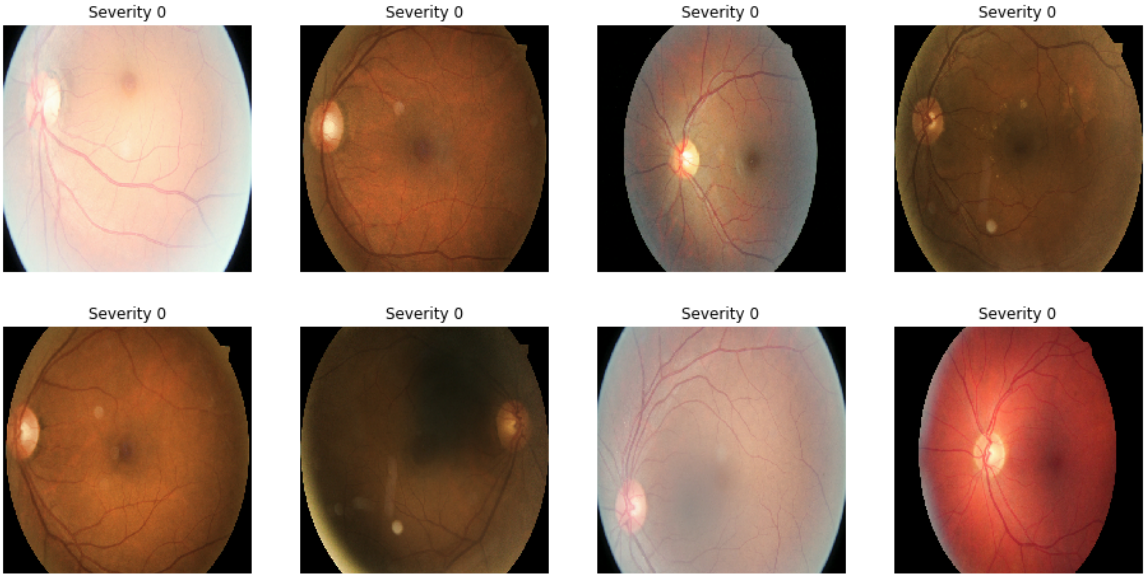


Eyepacs - 2:



Need for preprocessing of images:

Retino fundal photography can be done using various kinds of devices which often leads to severe inconsistencies in the quality of images taken, which leads to wrong diagnosis and sometimes even a retest because the image might not contain(although eyes do have them, the device was unable to capture these minute features which can be due to various factors such as bad lighting condition or device malfunction, etc.) the vessels, cotton wool spots, microaneurysms, etc. The other important factor due to which image processing is needed in order to highlight the cotton wool spots and the microaneurysms which act as definitive identifiers for a retinopathic eye. The following images are some examples of “bad” quality of images:



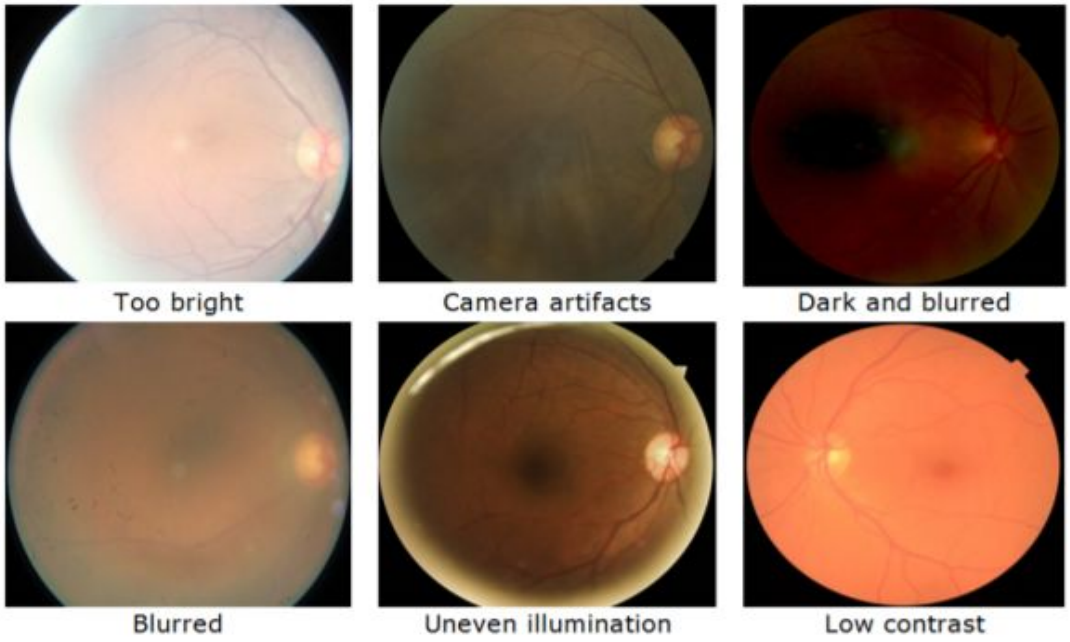
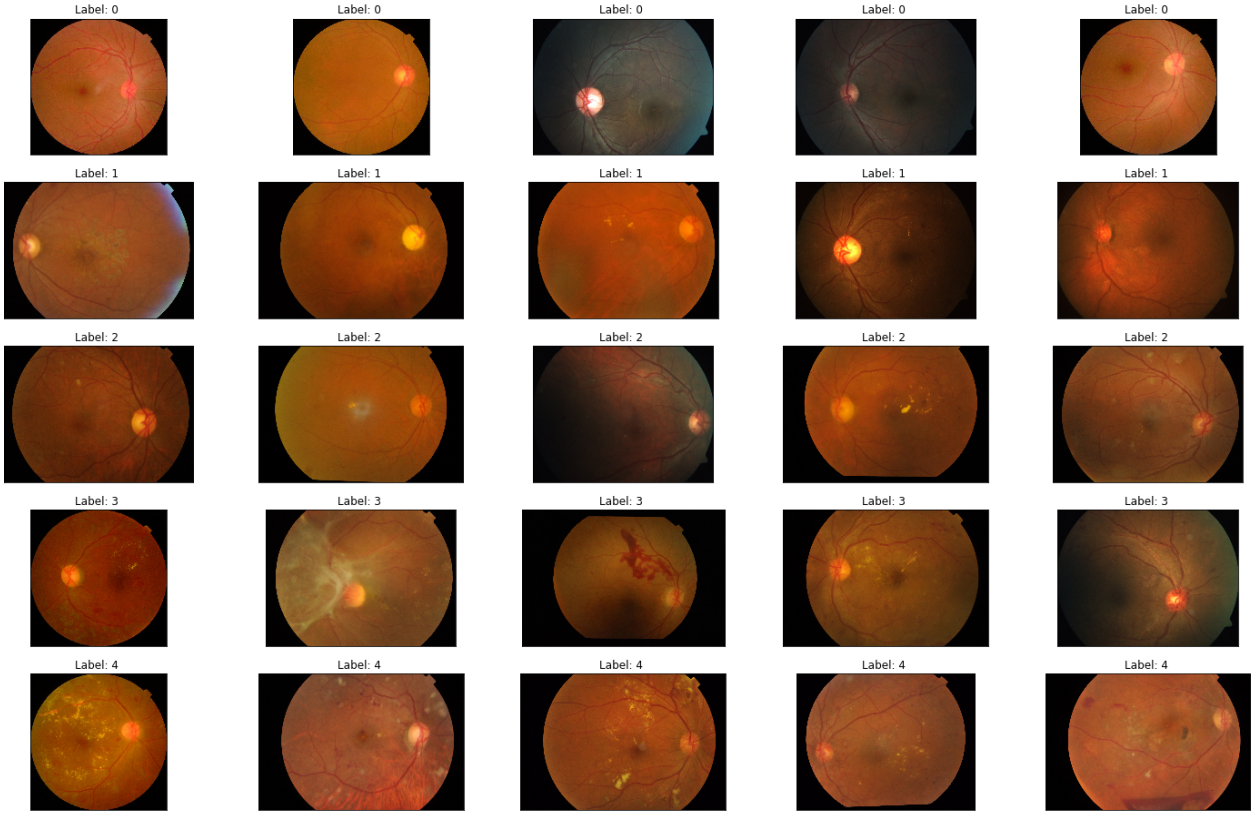


Figure 2: Retinal images of deficient quality

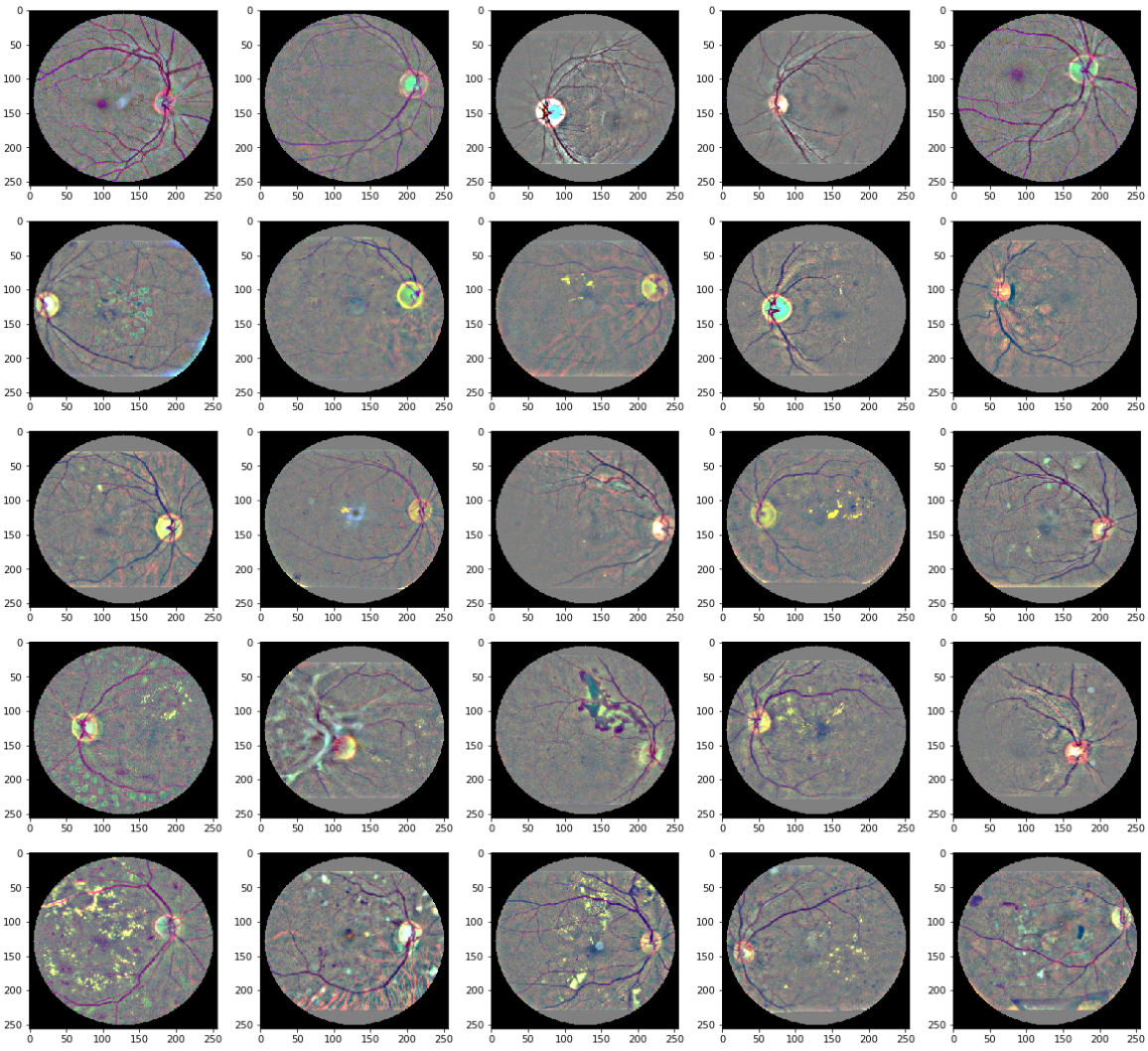
Image Preprocessing techniques used:

Median Subtraction: It is a common technique used to remove the background from images for semantic segmentation tasks and object recognition tasks.

Original unprocessed image:

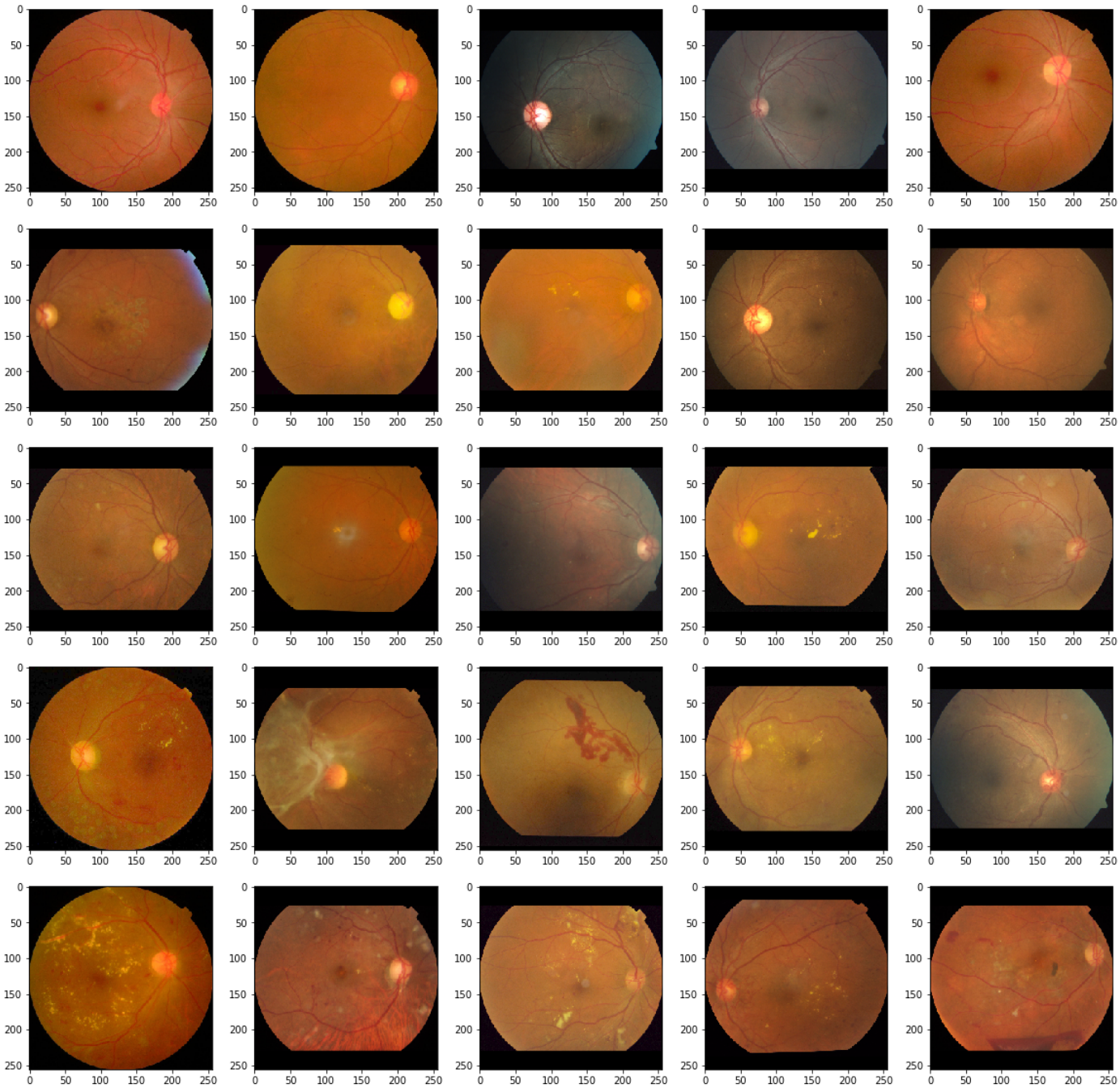


After applying median subtraction:



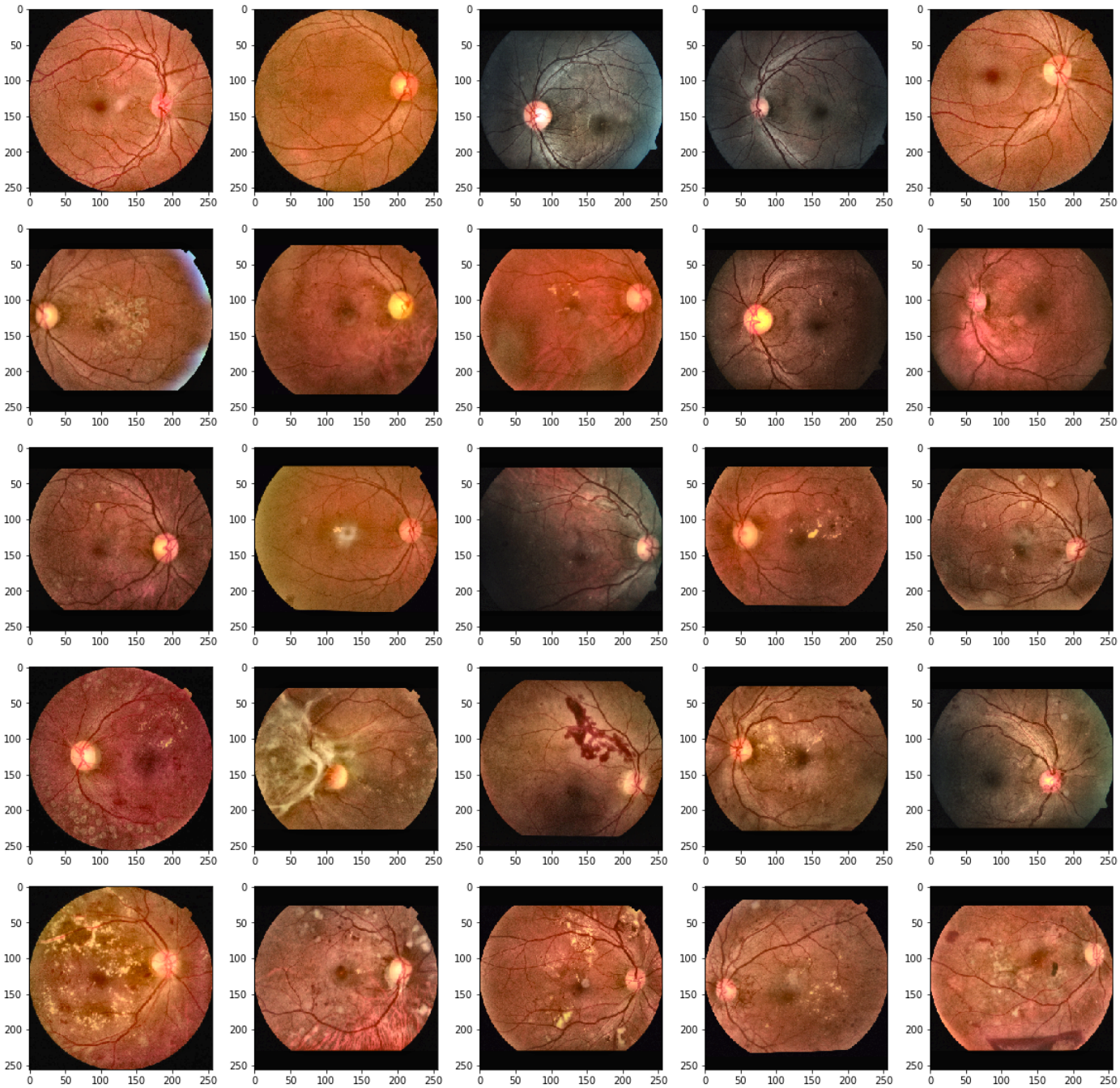
2)Gamma Correction-Gamma is an important but seldom understood characteristic of virtually all digital imaging systems. It defines the relationship between a pixel's numerical value and its actual luminance. Without gamma, shades captured by digital cameras wouldn't appear as they did to our eyes (on a standard monitor). It's also referred to as gamma correction, gamma encoding or gamma compression, but these all refer to a similar concept. Understanding how gamma works can improve one's exposure technique, in addition to helping one make the most of image editing.

After applying Gamma Correction:



3)Adaptive Histogram Equalization: Adaptive histogram equalization (AHE) is a computer image processing technique used to improve contrast in images. It differs from ordinary histogram equalization in the respect that the adaptive method computes several histograms, each corresponding to a distinct section of the image, and uses them to redistribute the lightness values of the image. It is therefore suitable for improving the local contrast and enhancing the definitions of edges in each region of an image.

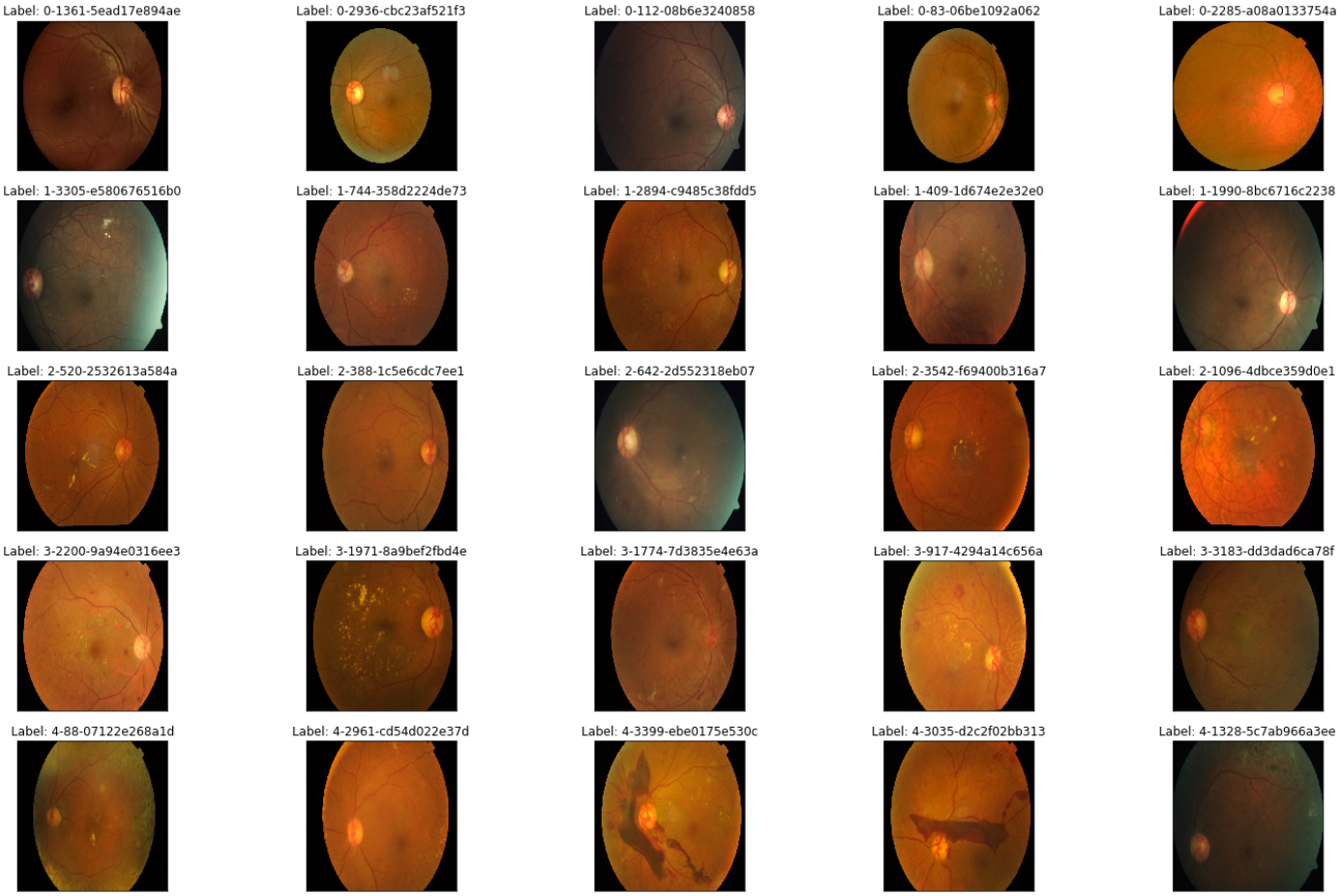
Results:



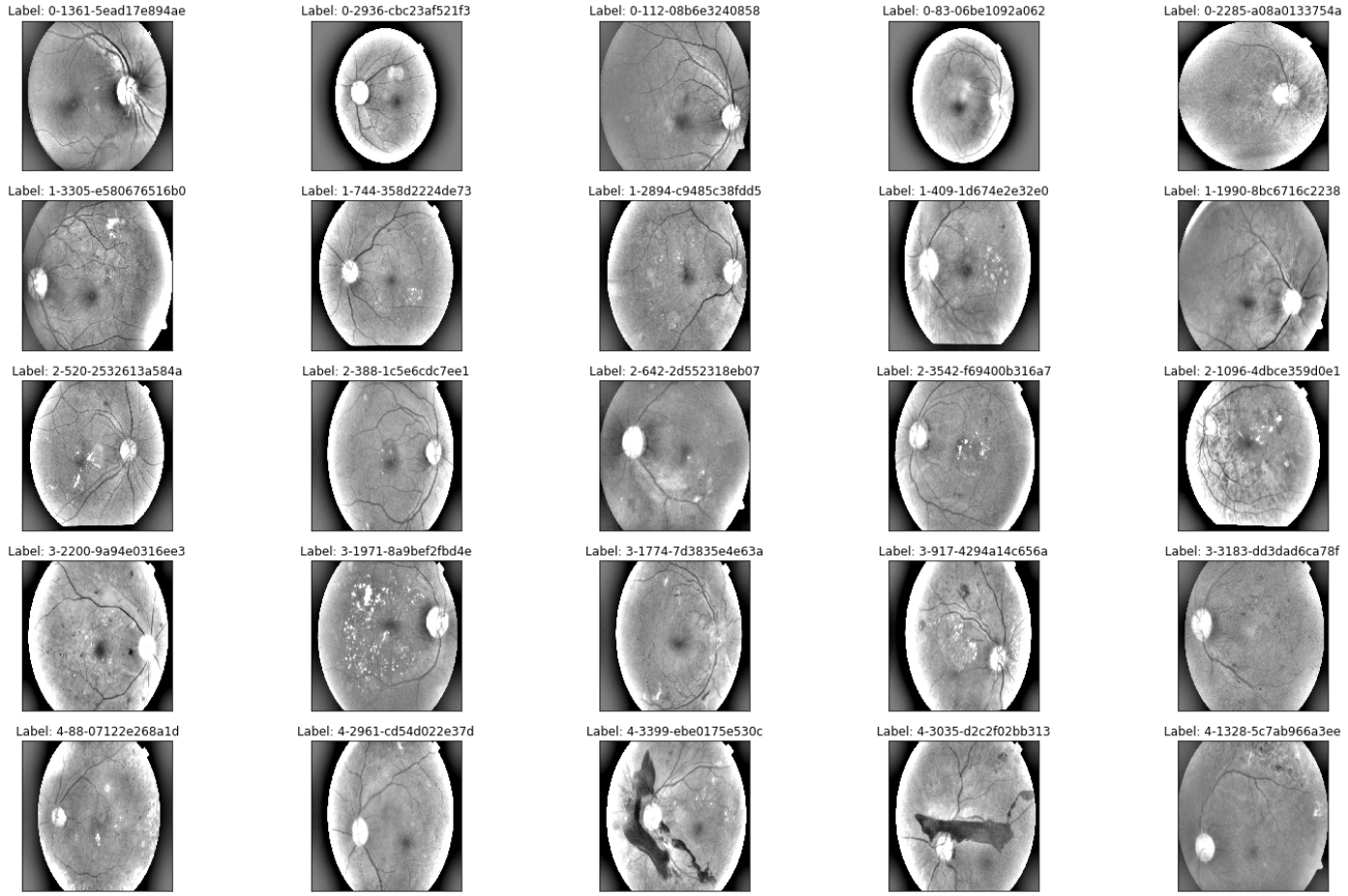
4)The key Image preprocessing technique in diabetic retinopathy detection is called as Ben Grahams's Image processing technique: In the last competition on DR detection at Kaggle, Ben Graham (last competition's winner) share the insightful way to improve lighting condition. Here, we apply his idea and can see many important details in the eyes much better. For full details, please refer to his technical report in the link above.

Ben Graham's preprocessing gives the best results because of how it highlights the cotton wool spots and the microaneurysms.

Results:(Original Image)



Preprocessed Image:



Some more results:

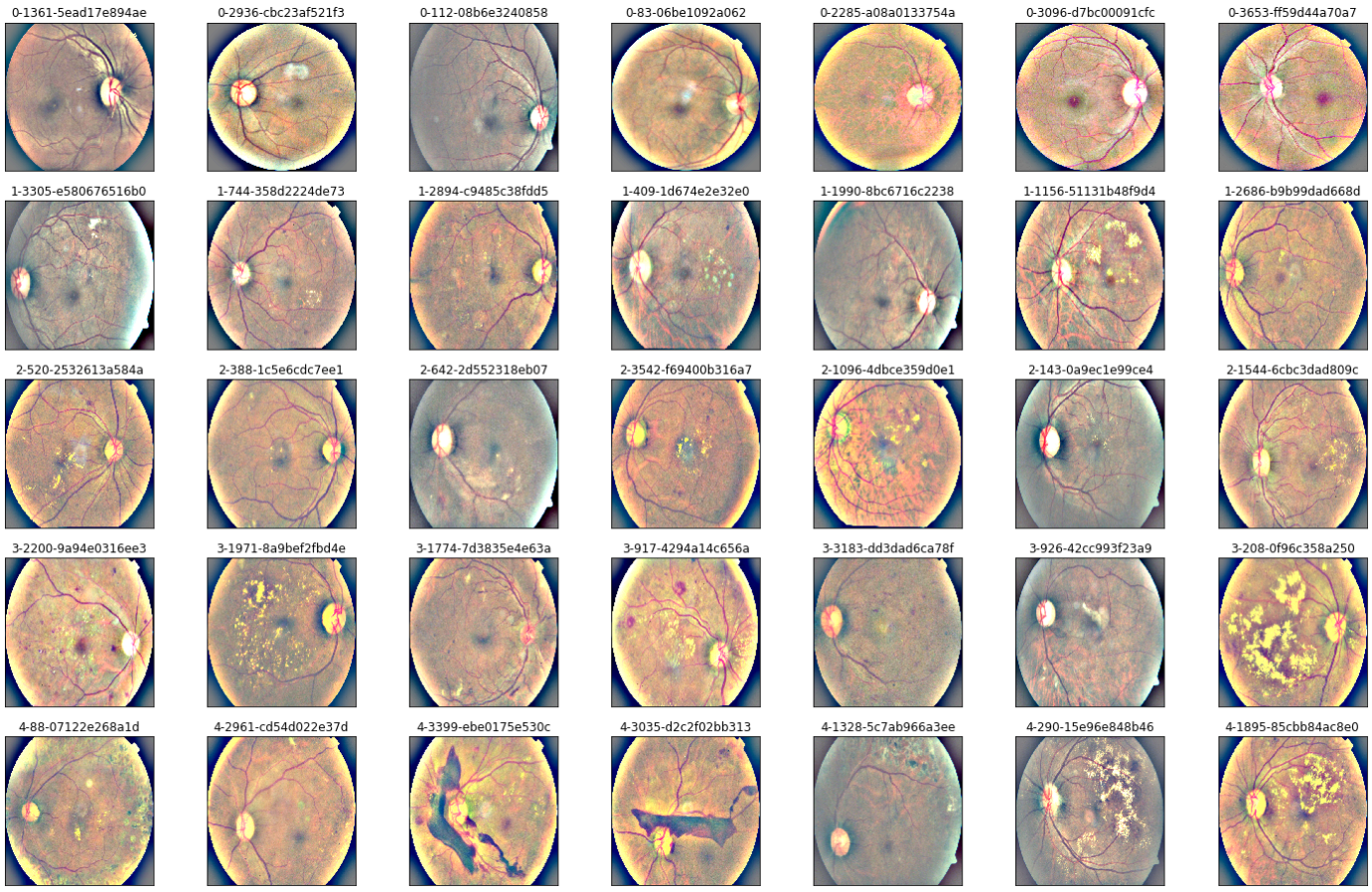
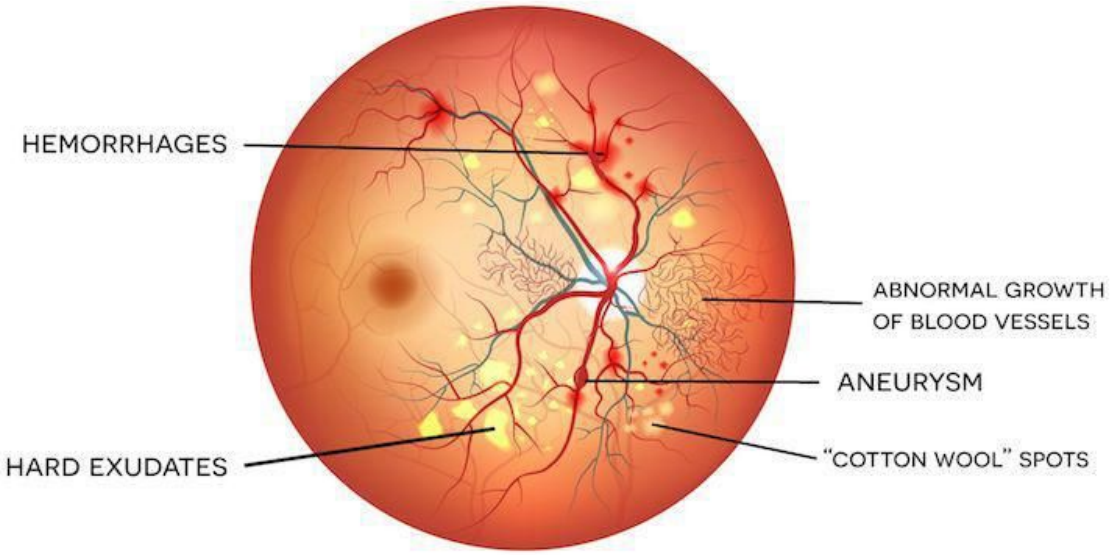
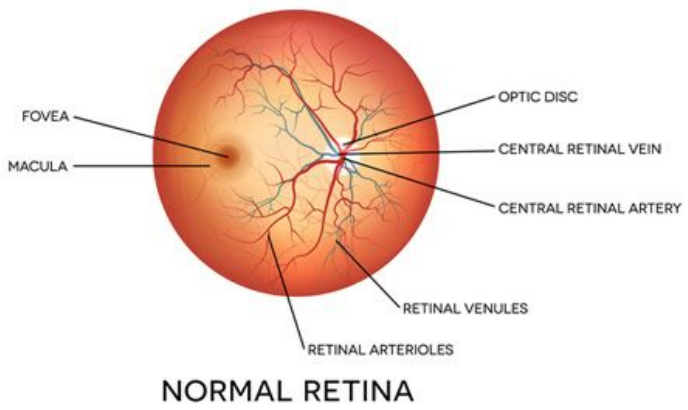


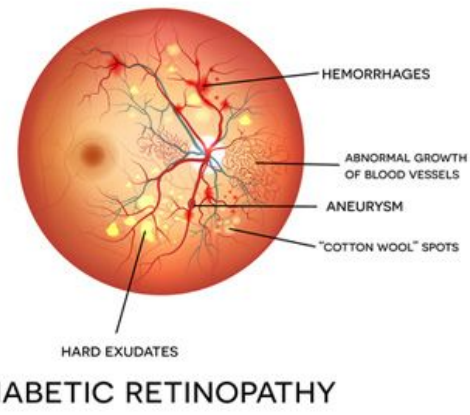
Image Preprocessing results:

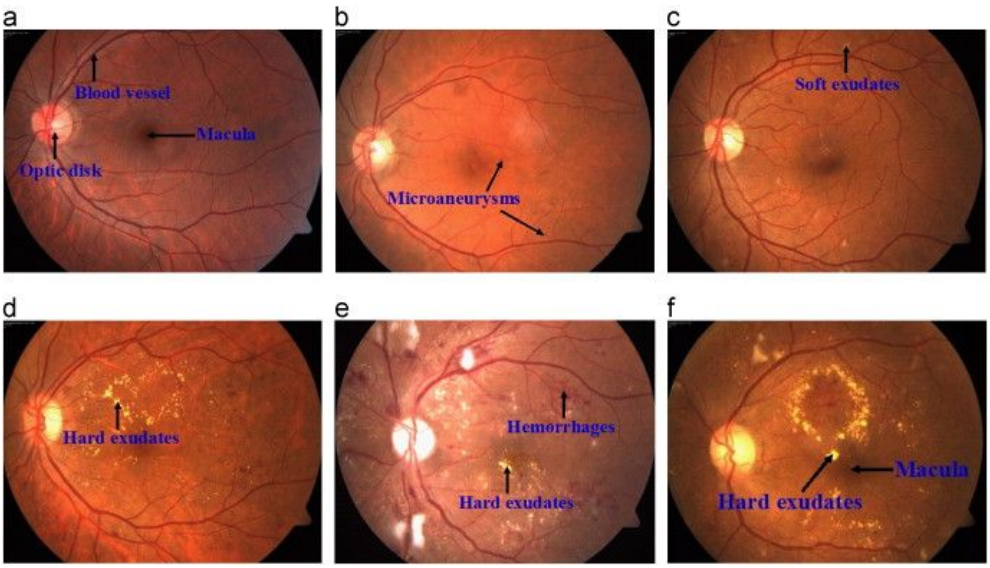
The best results were given by ben graham's image processing technique, hence we have utilized his technique throughout this project for image processing.

Diabetic retinopathy, what to detect?









Hard Exudates: Hard exudates (lipid) Hard exudates are small white or yellowish-white deposits with sharp margins. Often, they appear waxy, shiny, or glistening. They are located in the outer layers of the retina, deep to the retinal vessels.

Cotton wool spots(also called as soft-exudates) and Microaneurysms, the detectors for a retinopathic eye: Cotton wool spots are an abnormal finding on the funduscopic exam of the retina of the eye. They appear as fluffy white patches on the retina. They are caused by damage to nerve fibers and are a result of accumulations of axoplasmic material within the nerve fiber layer. There is reduced axonal transport (and hence backlog and accumulation of intracellular products) within the nerves because of the ischemia. This then causes the nerve fibers to be damaged by swelling in the surface layer of the retina. A 1981 analysis concluded that "in most instances, cotton-wool spots do not represent the whole area of the ischaemic inner retina but merely reflect the obstruction of axoplasmic flow in axons crossing into much larger ischaemic areas".[1] Associated findings include microvascular infarcts and hemorrhages. The appearance of cotton wool spots may decrease over time. Abundant cotton wool spots are seen in eyes affected by Diabetic retinopathy. The first stage, called non-proliferative diabetic retinopathy (NPDR), has no symptoms. Patients may not notice the signs and have 20/20 vision. The only way to detect NPDR is by fundus photography, in which microaneurysms (microscopic blood-filled bulges in the artery walls) can be seen. If there is reduced vision, fluorescein angiography can show narrowing or blocked retinal blood vessels clearly (lack of blood flow or retinal ischemia).

Now as we know what matters are the cotton wool spots and microaneurysms caused due to filling of blood which leads to bulging of the artery walls present in the retina.

The Transfer Learning Approach:

The aim of this study is to produce a Robust model for DR detection over the 5 classes to produce a model that is robust as well as is able to generalize well when faced with different kinds of images taken in different conditions with different kinds of devices.

All the works stated earlier are based on deep learning models and they generally use Precision, Recall, Specificity and sensitivity as metrics to measure and compare the performance of the produced model. In this study, our main focus will be to optimize AUC, and more specifically Quadratic Weighted Kappa or QWK. Quadratic Weighted Kappa or QWK has been one metric ignored by quite a few researchers because Quadratic Weighted Kappa or QWK obtained in DR detection usually is quite less around 0.83-0.88 and as this is a multi-class classification problem QWK is a metric that gives very accurate measure because simply of the fact how it calculates the score e.g if model predicts the class as 4 but the actual class is 3 penalties will be less in that proportion but if actual class is 0 and model predicts 4 we are way off and penalty on the QWK is proportional to this.

Thus QWK has now been made a standard by eyepacs and many organizations to measure the Robustness of the Models produced for the task of DR detection and has also been used as a metric to evaluate various other models giving a uniform scale to measure other model performance as well as its robustness and generalization capabilities.

For this task we have used the following models:

As it is already known transfer learning does give really good results when done sensibly, so our starting approach was to try out pre-trained models which included:

Resnet 50,EfficientNet B5,B6,B4,B7,Res-Next 32x4d.

Finally, we produced an ensemble of all the above models.

Understanding Quadratic Weighted Kappa(QWK):

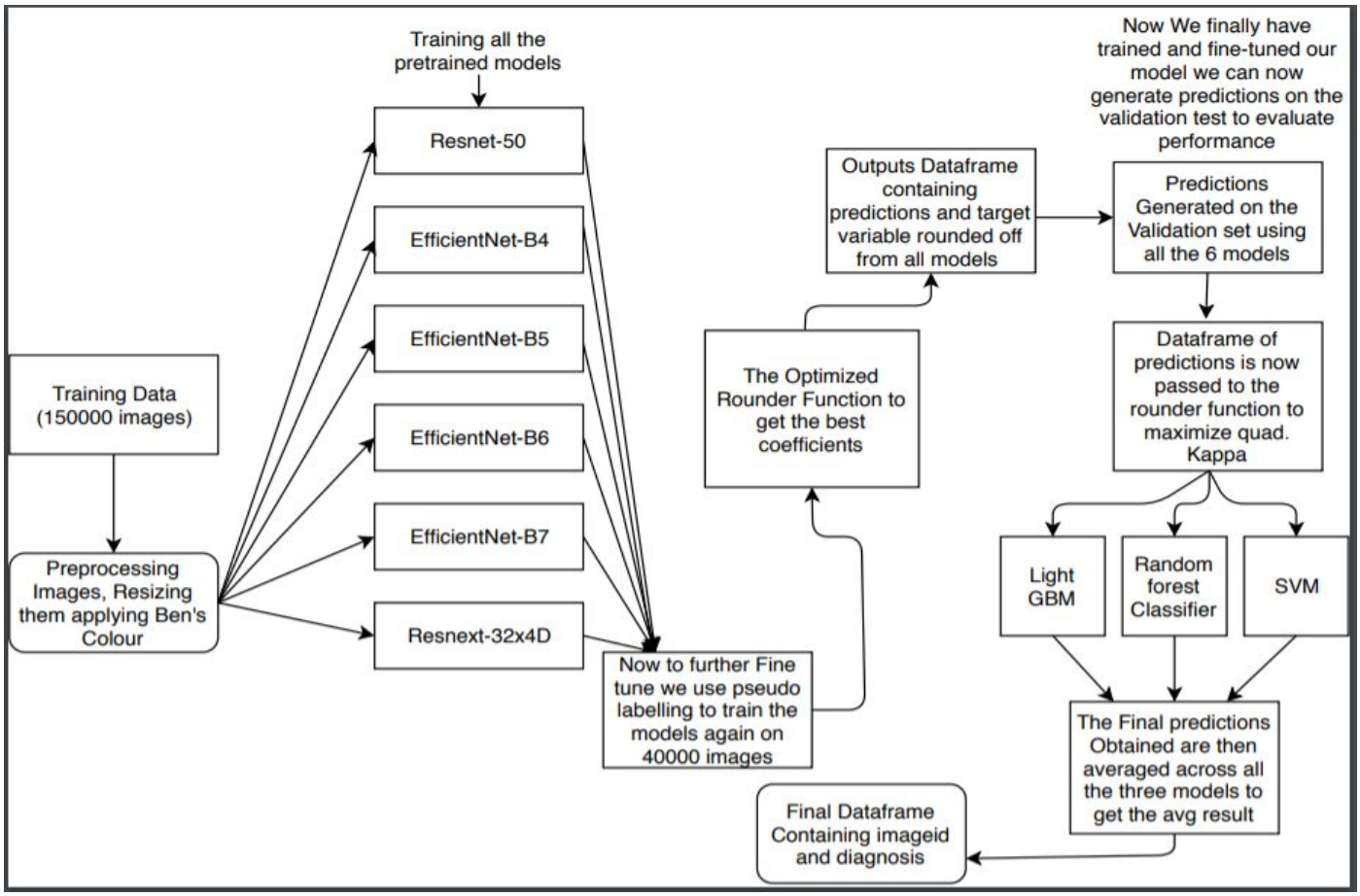
Kappa is the Chance adjusted index for the reliability of categorical measurements. This means that it accounts for the amount of agreement between raters that can be expected to have occurred due to chance (i.e., random guessing). This is unlike accuracy where you can get relatively fair score by intelligently random guessing. Here 3 matrices are involved:

1. N×N(N is the number of categories) histogram matrix O, where each element eij of O corresponds to the number of observations that received a category i by A and a category j by B. In our case, N=5 so O is 5×5 matrix. Each element eij will represent the count of images that received category i by A(say human) and category j by B(our models). So greater the number in diagonal, greater good.
2. N×N weights matrix W, where each element is calculated using the distance between ratings. More on this later.
3. N×N histogram matrix of expected rating E, which is calculated as the outer product between each rater's histogram vector of ratings. E is normalized so that E and O have the same sum. Now, each cell in O is multiplied by the corresponding cell in W and sum the results across all the cells. Call this Po. The same is done for E. Call this Pe. Then kappa is calculated as below:

$$\text{kappa}_{LW} = \frac{P_{observed} - P_{expected}}{1 - P_{expected}}$$

$$\text{weight} = 1 - \frac{(\text{distance})^2}{(\text{maximum possible distance})^2}$$

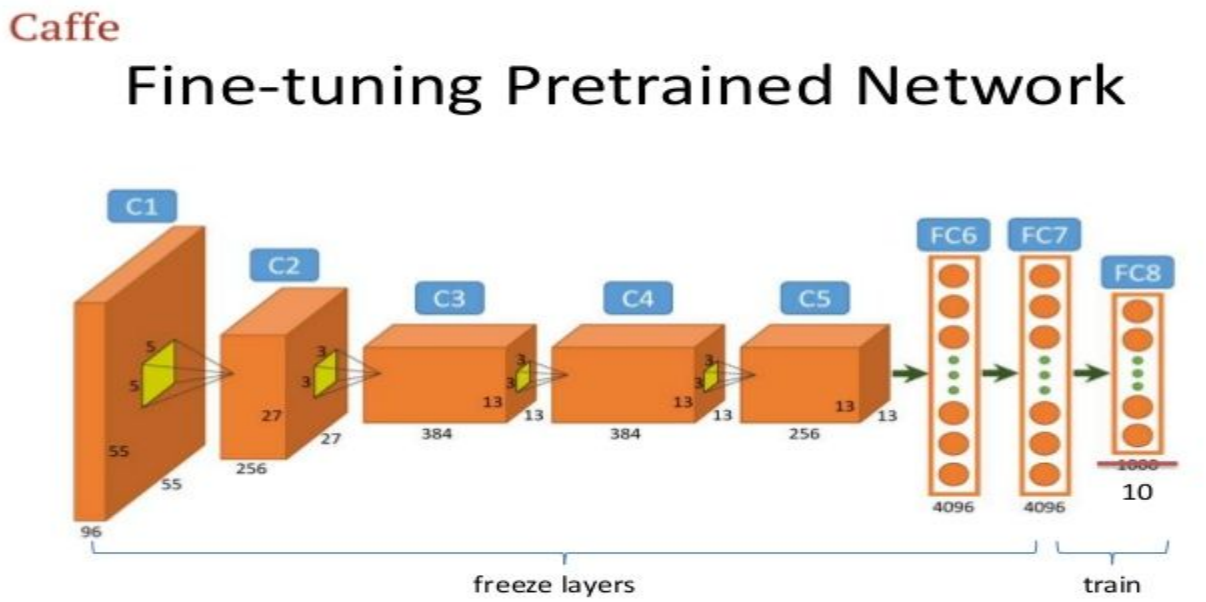
Proposed Model Framework:



We train our models on 1,50,000 images then further fine-tuned the models using 40000 images through a very popular unsupervised learning technique known as Pseud-Labeling. We validate our models on the remaining 15000 images. We finally test our models on the remaining 13,000 images using Test Time Augmentation.

Understanding the proposed framework:

- The training data consisting of 150000 images first goes through the preprocessing stage where the images are resized and ben graham's technique is applied.
- After the images have been resized and processed 5 pre-trained models already trained on imagenet are retrained on the 150000 retino-fundal photographs, but rather than just training the neck of the network which is the usual case when applying transfer learning, we unfroze all the backbones of all the networks hence essentially meaning we also re-train the convolution portion of the network on the images.



The convolution layers as it can be seen from the above diagram are generally kept frozen and only the 'Neck" of the network is trained. We unfroze the whole network and essentially reinforcing the learning of many subtle features that represents a retinopathic eye by training the convolution part to learn more about what kind of eye has diabetic retinopathy.

- After all the models were trained, they were further fine-tuned using an unsupervised learning technique known as pseudo labeling. So, pseudo- labeling a trick one can call it, but it is more than that. This is an excerpt from what Dr. Bengio said while referring to the increasing use of pseudo-labeling for improving model performance "*To climb the AI ladder with supervised learning may require "teaching" the computer all the concepts that matter to us by showing tons of examples where these concepts occur. This is not how humans learn: yes, thanks to the language we get some examples illustrating new named concepts that are given to us, but the bulk of what we observe does not come labeled, at least initially".*

First proposed by Lee et. al. in 2013 [1], the pseudo-labeling method uses a small set of labeled data along with a large amount of unlabeled data to improve a model's performance. The technique itself is incredibly simple and follows just 4 basic steps:

- Train model on a batch of labeled data.
- Use the trained model to predict labels on a batch of unlabeled data.
- Use the predicted labels to calculate the loss on unlabeled data.
- Combine labeled loss with unlabeled loss and backpropagate.
    -

And continue repeating the process, but there is a very small concept on why this works well, and let's see how:

Pseudo-labeling trains the network with labeled and unlabeled data simultaneously in each batch. This means for each batch of labeled and unlabeled data, the training loop does:

- One single forward pass on the labeled batch to calculate the loss → This is the labeled loss
- One forward pass on the unlabeled batch to predict the "pseudo labels" for the unlabeled batch
- Use this "pseudo label" to calculate the unlabeled loss.
- Now instead of simply adding the unlabeled loss with the labeled loss, Lee proposes using weights. The overall loss function equation looks like this:

$$L = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{C} L(y_i^m, f_i^m) + \alpha(t)\frac{1}{n'}\sum_{m=1}^{n'}\sum_{i=1}^{C} L(y_i'^m, f_i'^m), \tag{15}$$

$$\text{Loss per Batch} = \text{Labeled Loss} + Weight * \text{Unlabeled Loss}$$

Here the loss function L represents categorical-Cross Entropy which is given as follows:

$$C = -\frac{1}{n}\sum_{x}\sum_{j}\left[y_j \ln a_j^L + (1-y_j)\ln(1-a_j^L)\right].$$

In equation(15), the weight (alpha) is used to control the contribution of unlabeled data to the overall loss. In addition, the weight is a function of time (epochs) and is slowly increased during training. This allows the model to focus more on the labeled data initially when the performance of the classifier can be bad. As the model's performance increases over time (epochs), the weight increases and the unlabeled loss has more emphasis on the overall loss.

After all the models have been fine-tuned, we further improve the QWK score for every model using an optimized rounder function. We treat this problem as a regression problem rather than treating it as a classification problem, this is because the model performance was comparatively far better when using regression rather than using classification. The table below represents the results of two models trained on 150000 images and then validated across 8000 images after completing pseudo-labeling.

Model used	Training QWK for classification.	Testing QWK for classification.	Training QWK for Regression. (using optimized rounder)	Testing QWK for regression. (using optimized rounder)
EfficientNet-B5	0.8752	0.863	0.9181	0.9021
Resnet-50	0.8522	0.8411	0.90451	0.904973

The above results clearly show the regression model outperforms classification models, we further utilized an optimized rounder function, which takes in the predicted results and knows the actual results and calculates the optimal co-efficient values for each class.

2nd level of Ensembles using Machine Learning:

After all the models were trained we further predicted the results using all those models in our case we have 6 models. After that, we created a data frame with 7 columns the first 6 columns contained predictions from the 6 models that we earlier trained, the last column represents the true-label for each image.

- The created data frame then was sent to three machine learning algorithms to find hidden patterns in our model prediction this further boosted our average QWK score from 0.9352 to 0.9417.
- The three models used are - lightGBM, Random Forest Classifier, and SVM.
- After the results were obtained from these models, the predictions were then simply averaged to get the final prediction i.e. ImageID and Diagnosis.

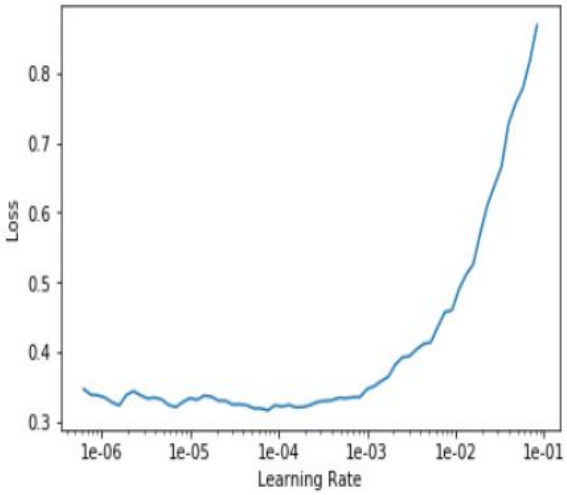
Training procedure, Results, and Visualizations:

- We Train all the Pretrained models again on 150000 images by unfreezing the weights in the backbone layers of the Network.
- **We Trained the following models:**
- Efficient net B7,B6,B5,B4.
- Resnet 50.
- Resnext 32x4d.
- **For all the models Following Data Augmentations were used:**
- do_flip
- flip_vert
- Max_zoom.
- Rotate-60,90,120.
- Differential Learning rates were used using FAST.AI API and its Learner function.
- We used RAdam as the Optimizer.
- Batch Size for Efficient Nets was kept at -84 and for Resnets-120 and for resnext-124.
- All the Images were Resized to 3 image categories as part of the Technique called

Progressive resizing as follows:

- 240,360,440
- All the models were each trained on images with 240 the 360 and 440 thus in total 3 models for each model hence in total 18 models . 1 model is trained across all the three image sizes to complete its training.
- For Test-Time Augmentations we used the following data augmentations:
- do_flip
- flip_vert
- Rotate-60,90.

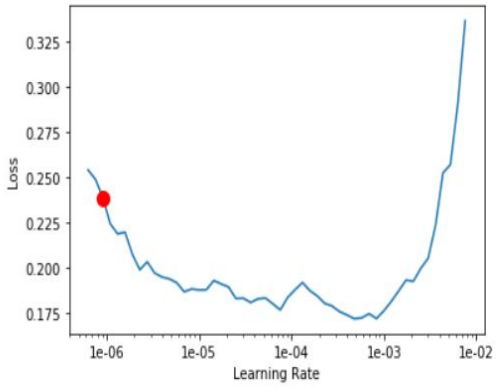
Differential learning rates were used using the "learner" function provided by the fast.ai API, below are the graphs of how learning rate vs Loss varies, which gives one idea about the range in which the learning rate should be kept in order to support maximal learning of all the models.



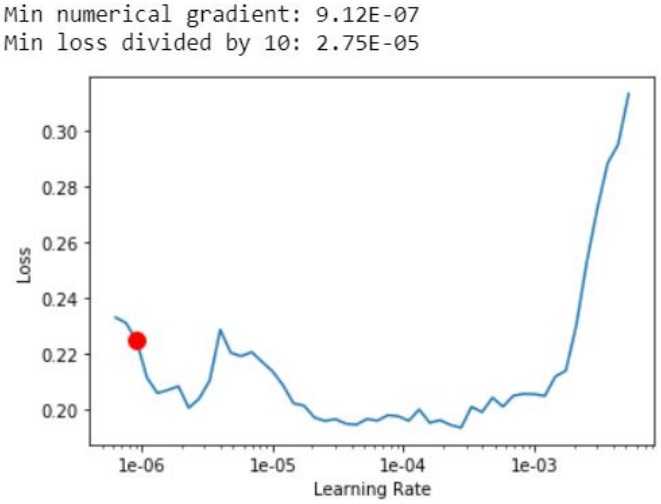
Learning rate vs Loss value for Resnet-50.

Here we can see that the value of loss function falls rapidly in the range 1e-01 to 1e-04. Hence this range can now be used to vary the learning rate and initiate maximum learning for the model. This is the most common and intuitive way to interpret these graphs, the other graphs are the same and can be interpreted in the same way.

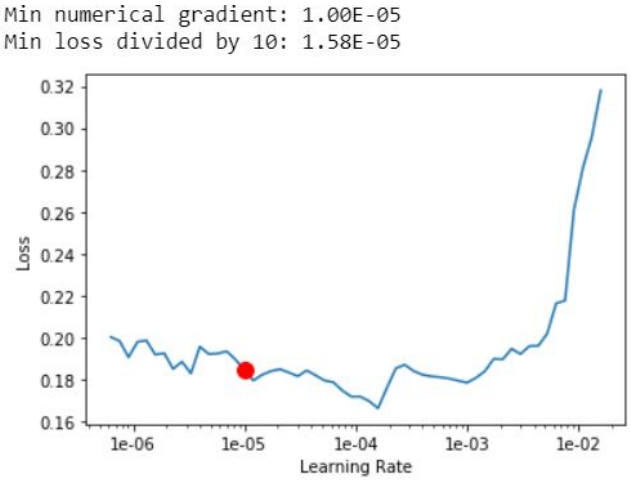
```
LR Finder is complete, type {learner_name}.recorder.plot() to see the graph.
Min numerical gradient: 9.12E-07
Min loss divided by 10: 4.79E-05
```

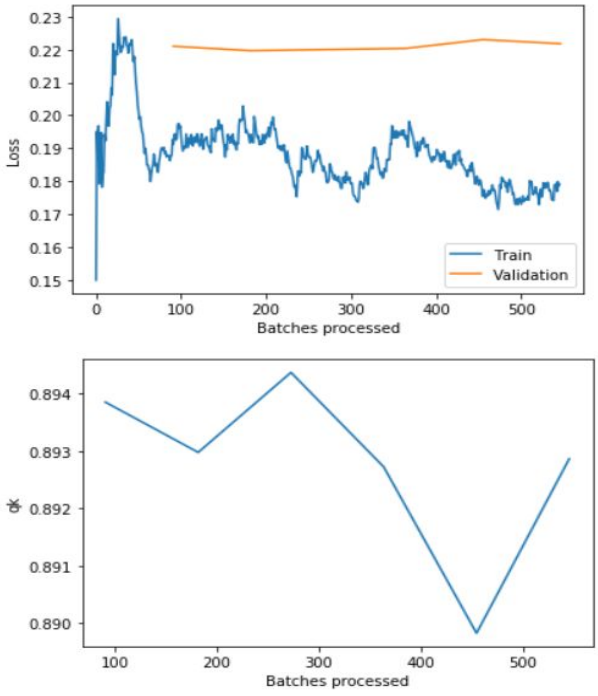


Learning rate vs loss graph for EfficientNet-B4.



Learning rate vs Loss graph for Efficientnet-B5.



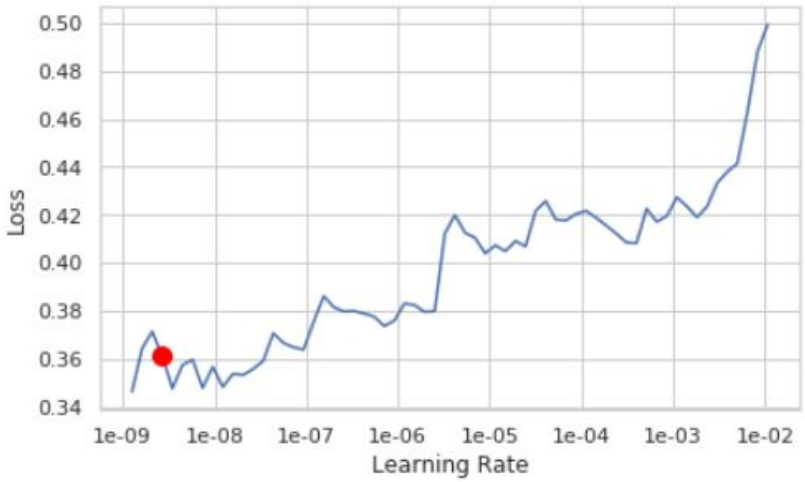


(a)

(a)Learning rate vs Loss graph for EfficientNet B6.

(b)Loss value for validation and testing,

And QWK score vs total batches processed. (b)



Learning rate vs loss graph for Res-Next32x4d.

Results:

Model Architecture	Training QWK.	Testing QWK
Resnet-50.	0.90451	0.904973
EfficientNet-B4	0.914032	0.918952
EfficientNet-B5	0.910874	0.920012
EfficientNet-B6	0.911269	0.912452
EfficientNet-B7	0.925656	0.911023
Res-Next 32x4d	0.895241	0.901085

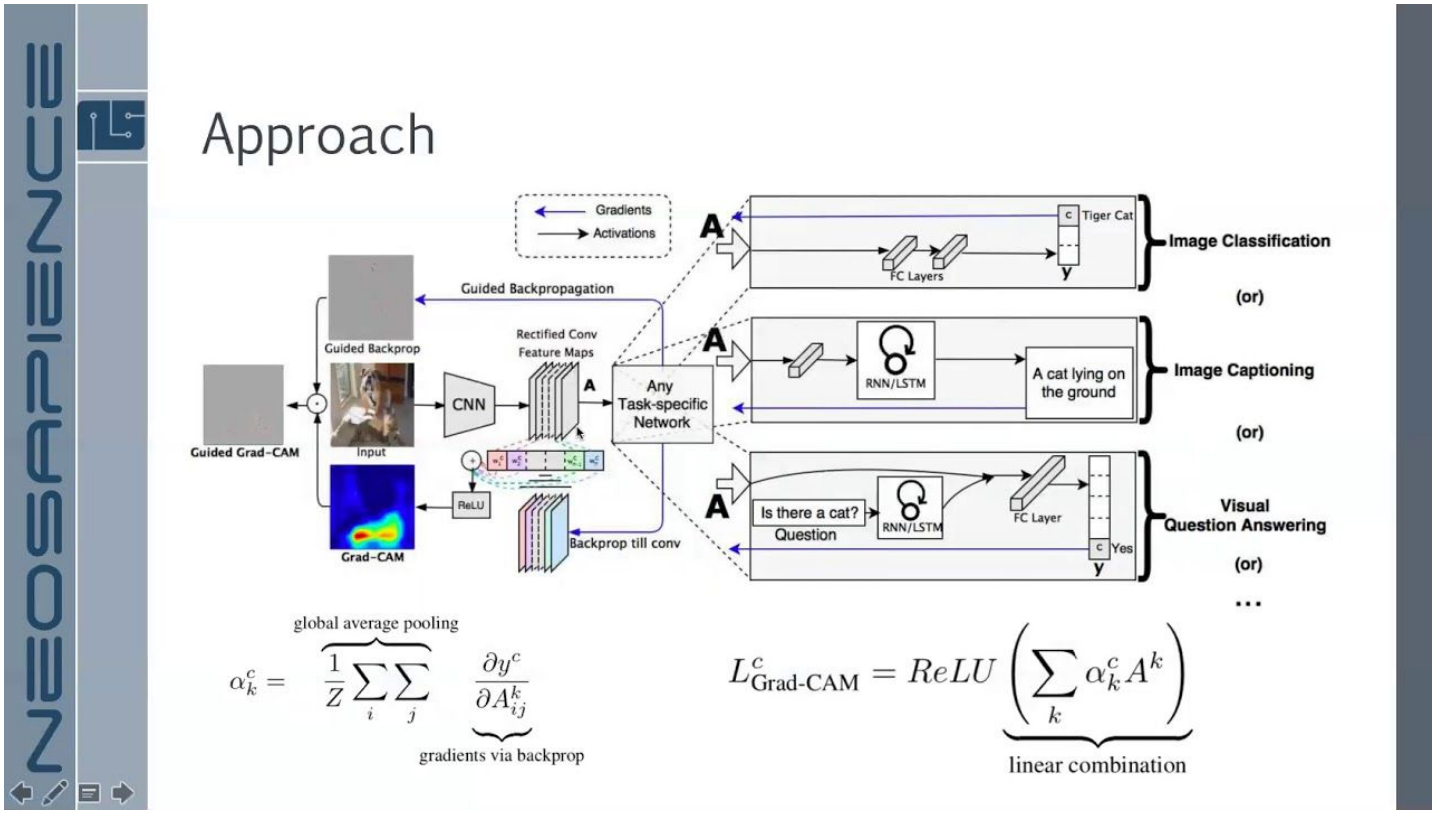
These Individual results when averaged using a weighted voting scheme following the rules of convex optimization the **Training QWK was 0.930021 and Testing QWK was = 0.935216.**

Final Results comparison:

S.NO	Authors	Year	Technique	QWK	AUC-ROC Metric	Dataset Used
1	Varun Gulshan, Ph.D.; Lily Peng, MD, Ph.D.; Marc Coram, Ph.D.; Martin C. Stumpe, Ph.D.; Derek Wu, BS; A	2016	Inception V-3.+image processing.	0.8851(5)	0.991(3 vs All), 0.993(Avg-AUC across all datasets)	Eyepacs-1(15,000, Messidor-2(2000), Google's own DB-1,20,000, APTOS(13000)
2	Daniel Shu Wei Ting, MD, PhD[1,2]; Carol Yim-Lui Cheung, PhD[1,3]; Gilbert Lim, PhD[4]; et al	2017	Inception V-3 and V-4 + CLAHE	0.856(5)	0.971(4 vs All), Avg AUC-0.951	Eyepacs-2(76000),Messidor-2(200)
3	Bellemo, Valentina, et al.	2019	VGG-net	0.831(3)	0.973(2vs alll) and	Eyepacs-2(71000),and Zambian eye society(18000)
4	Wei Zhang, Jie Zhong, Shijun Yang, Zhentao Gao	2019	Two-Part ensemble. 1-binary classification then, 2- grading System. 1-Xception,Resnet-50,Inception resnet-v2.	0.8771(4)	0.981(Classification), 0.994(Avg-AUC)	Sichuan Medical Center(15000)
5	Khwaja Wisal Maqsood and Manoj.	2019	EfficientNet+Resnet+Resnext+SVM+LGBM+RF+Image Processing+Pseudo labelling.	0.9417(5)	0.9971(4 vs all) 0.9959(avg AUC)	Eyepacs 1(120000) Eyepacs 2(84000) Messidor-2(2000) APTOS(8000)

Visualizing Convolutional Neural Networks

Results we obtained do break past results but what is our model actually seeing. Interpreting and analyzing the model is one of the most important aspects of creating a robust model that generalizes well because we need to understand where our model is failing and why. So for this, we use a technique called GRAD-CAM(46) which was engineered by researchers to understand the activation values of feature maps in the last Convolutional layer just before Pooling is applied.

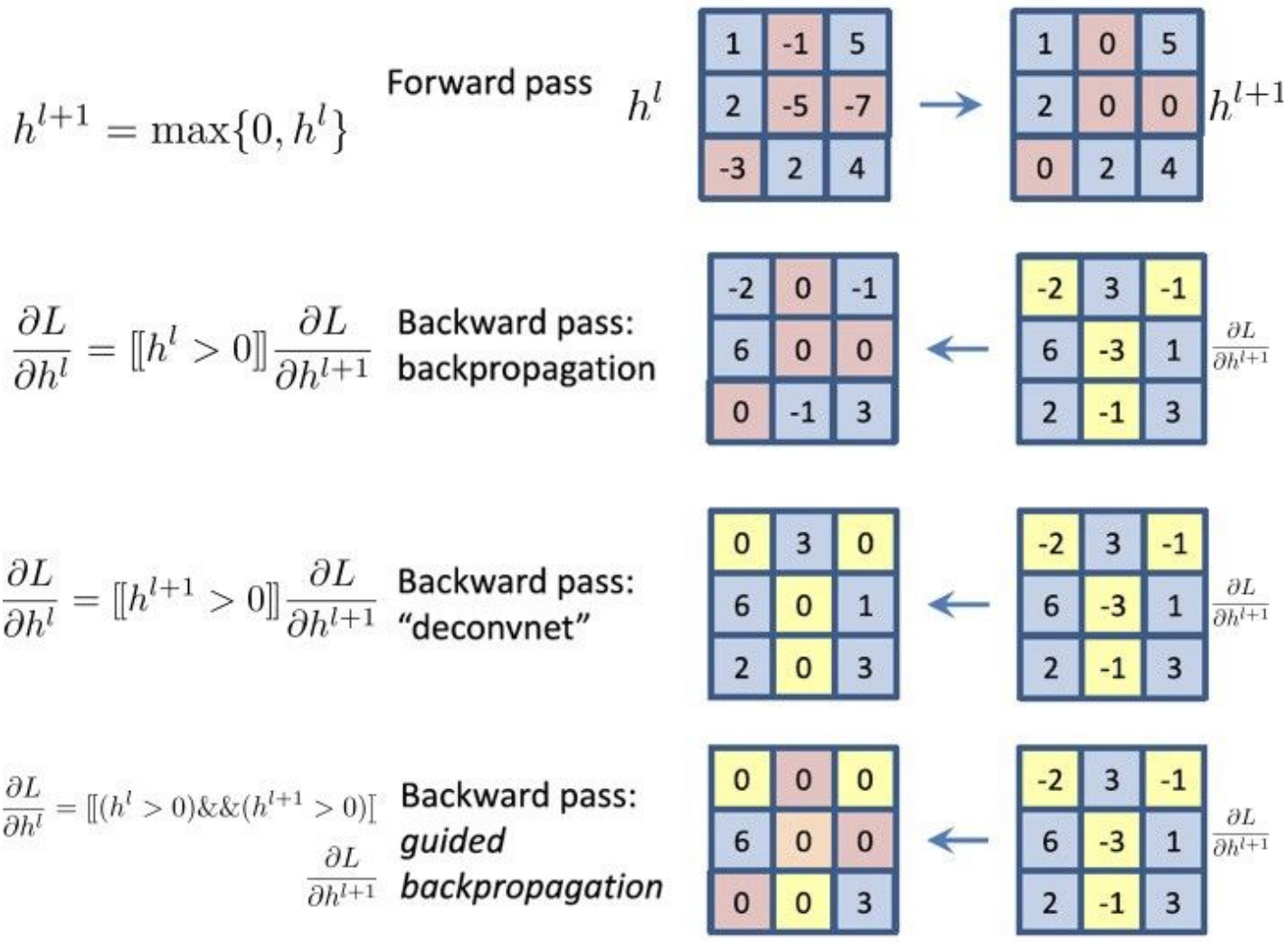


Basic idea behind GRAD-CAM:

- We believe that most important spatial information comes from the 3D-tensor of the *last convolutional layer* (just before the GlobalPooling layer), which is the nearest spatial information flowing to the last FC layer.
- For each channel of this 3D-tensor, each activated pixel region represents important features (e.g. blood vessel/scab/cotton wool) of the input image. Note that some features are important to determine class 0 (perfectly fine blood vessel), some features are important to determine class 4 (big cotton wools). Normally, we expect each channel to capture a different set of features
- To emphasize features that finally affected the final prediction, we calculate the **gradient of the final predicted class with respect to each feature.** If that feature is

important to this class, it should have a high gradient (i.e. increase the value of this feature, the prediction confidence increases)

- Therefore, we multiply the activated values of this 3D-tensor and gradients together, to obtain the visualized heatmap for each channel. Note that we have multi-channel, and each channel usually has multiple-features.
- Finally, we combine heatmaps of all channels using a simple average and remove negative value (the ReLu step in the above picture) to obtain the final heatmap.

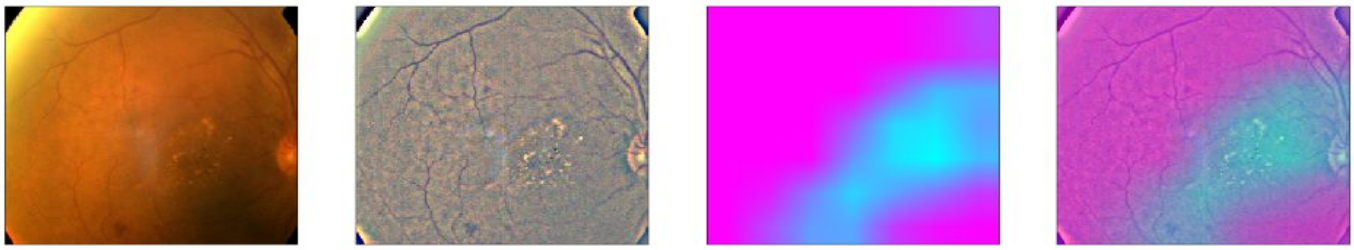


While Grad-CAM visualizations are class-discriminative and localize relevant image regions well, they lack the ability to show fine-grained importance like pixel-space gradient visualization methods (Guided Backpropagation and Deconvolution). For example, take the case of the left image in the above figure, Grad-CAM can easily localize the cat region; however, it is unclear from the low-resolutions of the heat-map why the network predicts this particular instance is 'tiger cat'. In order to combine the best aspects of both, we can fuse Guided Backpropagation and the Grad-CAM visualizations via a pointwise multiplication. The above figure illustrates what guided means and how the activation heat map is

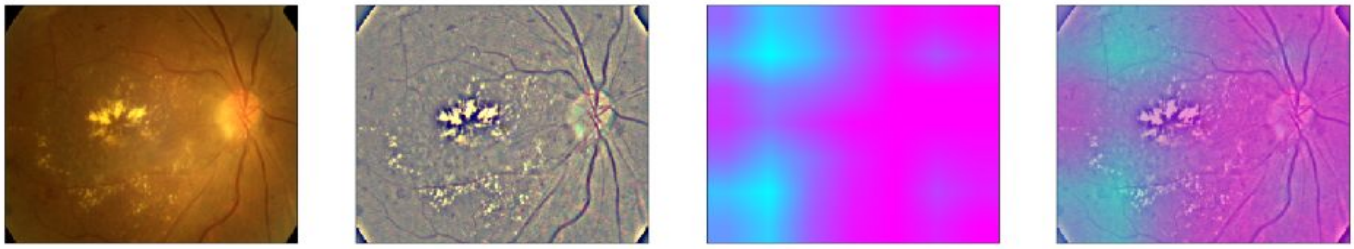
superimposed with guided activations and the original input image

Let’s see how our models performed:

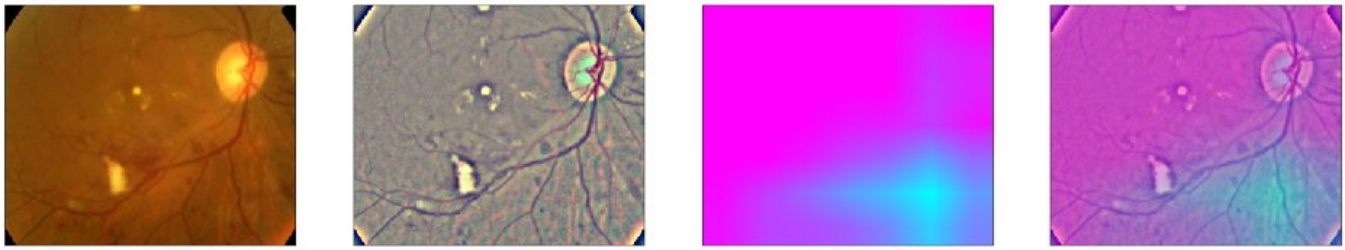
```
test pic no.1
raw output from model :
1.000 0.966 0.520 0.268 0.109
```



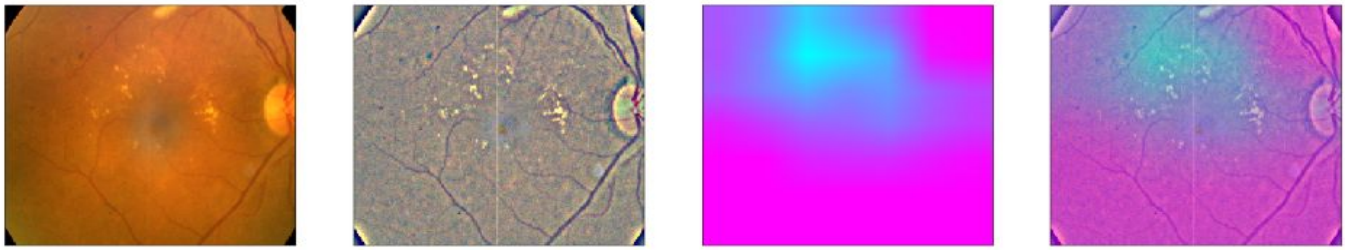
```
test pic no.2
raw output from model :
0.998 0.999 0.999 0.931 0.153
```



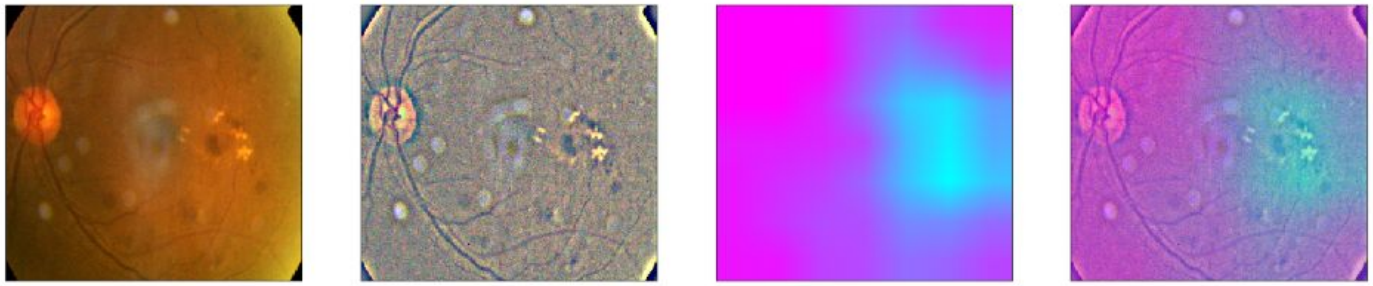
```
test pic no.3
raw output from model :
0.999 0.998 0.991 0.812 0.076
```



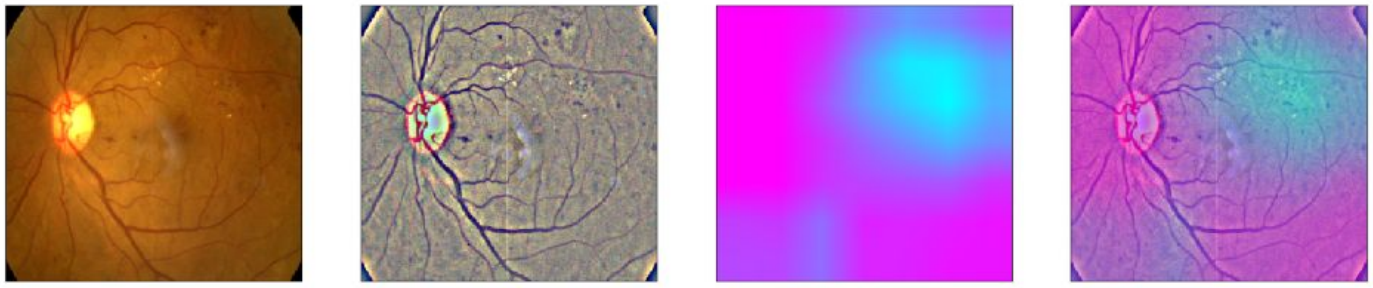
```
test pic no.4
raw output from model :
0.998 0.998 0.758 0.031 0.017
```



The above outputs are generated using Resnet-50.



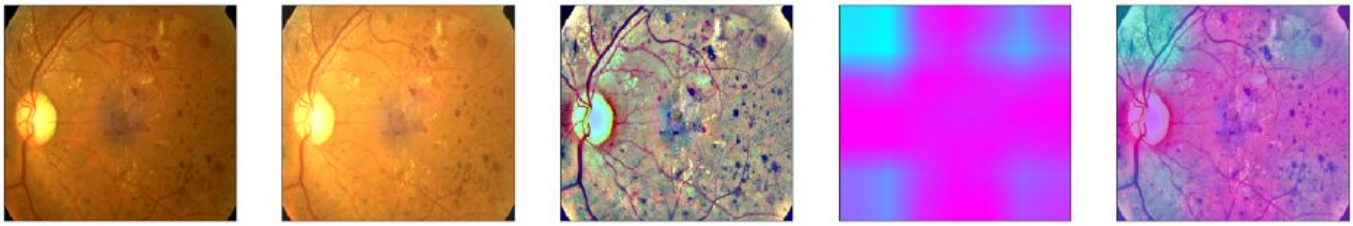
```
test pic no.6
raw output from model :
0.994 0.995 0.959 0.172 0.042
```



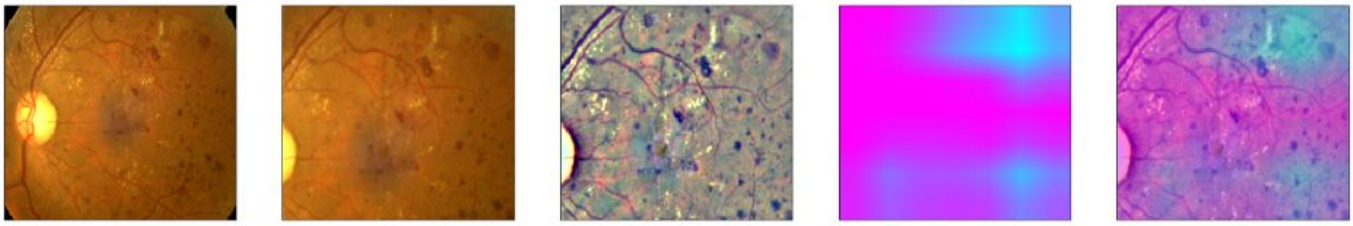
The above are visualizations obtained using EfficientNet.

Adding data augmentations:

```
test pic no.2 -- augmentation: brightness or contrast
raw output from model :
0.998 1.000 1.000 0.946 0.085
```

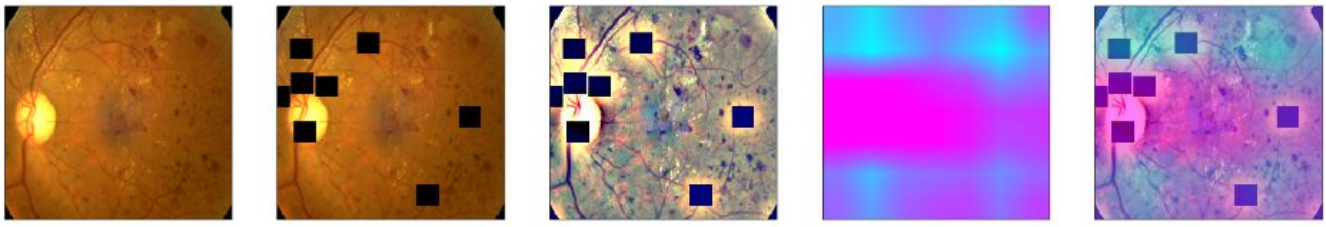


```
test pic no.3 -- augmentation: crop and resized
raw output from model :
0.997 1.000 0.999 0.897 0.063
```

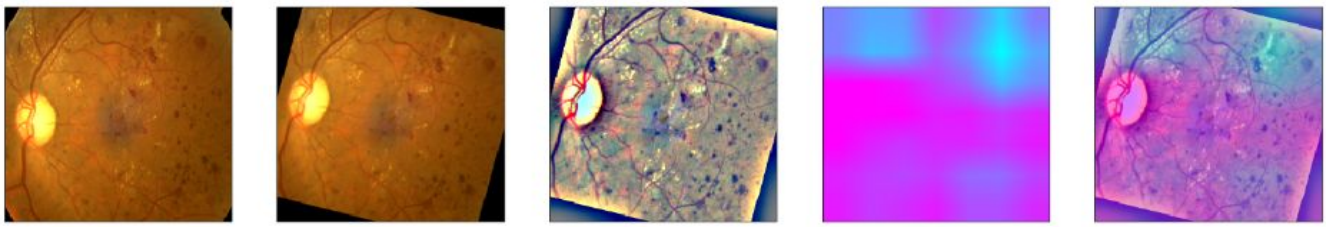


Some more Augmentations:

```
test pic no.4 -- augmentation: CutOut
raw output from model :
0.996 1.000 1.000 0.974 0.094
```



```
test pic no.5 -- augmentation: rotate or flip
raw output from model :
0.998 1.000 1.000 0.979 0.095
```



Conclusion and Future Work:

While these results are encouraging, there are a few limitations to this study. First, although we used the 5-point ICDR grade for training, the original and main algorithm used in this study only made a binary call in terms of DR. This is consistent with other deep-learning systems, including the IDx algorithm, that defines referable DR as moderate or worse DR and/or DME.14,15,23 In addition to the binary call, we also validated another model that returns a 5-point grade with slightly better performance than the algorithm used in this study (eFigure 1 in the Supplement), but this validation was performed retrospectively. The more granular 5-point grading would be especially helpful for screening programs in which patient treatment varies at each level of severity. In particular, the threshold for and timing of referrals in DR screening programs often depend on the resources of the program. Currently, in most programs, sight-threatening cases will be referred urgently while moderate cases without DME will be followed clinically. Identifying mild cases may also be of clinical value for health care systems that have a different screening interval for patients with no disease and mild disease.24 A more granular grading output, such as the 5-point ICDR scale, would also be more robust to guideline changes.

Although in this work we have produced a robust model that is able to generalize well, there are significant shortcomings, we still need to make deep learning architectures rely more on less data and more abstraction power. Thus, this does call upon the fact that we need some new revolutionary changes in how convolutional neural networks work and produce results, interpretability techniques such as Grad Cam has definitely helped in improving model performance we still need better architectural improvements.

Future Work:

While there are many avenues for future work, this study demonstrates the feasibility of using an automated DR grading system in health care systems and shows that the trained algorithm generalizes to this prospective population provided by the datasets.

References:

1. Zhang X, Saaddine JB, Chou C-F, et al. Prevalence of diabetic retinopathy in the United States, 2005-2008.JAMA. 2010;304(6):649-656.
2. Raman R, Rani PK, Reddi Rachepalle S, et al. Prevalence of diabetic retinopathy in India: Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study report 2. Ophthalmology. 2009;116(2):311-318.
3. Chakrabarti R, Harper CA, Keeffe JE. Diabetic retinopathy management guidelines. Expert Rev Ophthalmol. 2012;7(5):417-439.
4. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. 2013;131 (3):351-357.
5. Mookie MRK, Acharya UR, Chua CK, Lim CM, Ng EYK, Laude A. Computer-aided diagnosis of diabetic retinopathy: a review. Comput Biol Med. 2013;43(12):2136-2155.

6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444.
7.Classification of diabetic retinopathy and diabetic macular edema

Lihteh Wu, Priscilla Fernandez-Loaiza, Johanna Sauma, Erick Hernandez-Bogantes, and Marissé Masis.

8.An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness.
9.Bellemo, Valentina, et al. "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study."
11.Automated identification and grading system of diabetic retinopathy using deep neural networks Wei Zhang, Jie Zhong, Shijun Yang, Zhentao Gao, Junjie Hu, Yuanyuan Chen, Zhang Yi.
12Perez, Luis, and Jason Wang. "The effectiveness of data augmentation in image classification using deep learning." *arXiv preprint arXiv:1712.04621* (2017).
13.Gangnon, R. E., Davis, M. D., Hubbard, L. D., Aiello, L. M., Chew, E. Y., Ferris, F. L., & Fisher, M. R. (2008). A severity scale for diabetic macular edema developed from ETDRS data. *Investigative ophthalmology & visual science*, *49*(11), 5041-5047.
14.Gulshan, Varun, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *Jama* 316.22 (2016): 2402-2410.
15.Gargeya, Rishab, and Theodore Leng. "Automated identification of diabetic retinopathy using deep learning." *Ophthalmology* 124.7 (2017): 962-969.
16.Ting, Daniel Shu Wei, et al. "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes." *Jama* 318.22 (2017): 2211-2223.
17.Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." *Cell* 172.5 (2018): 1122-1131.
18. Cuadros J, Bresnick G. EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening. Journal of diabetes science and technology (Online). 2009;3(3):509-516 [attached]."
19. Chan J. C., Malik V., Jia W., Kadowaki T., Yajnik C. S., Yoon K.-H., Hu F. B. Diabetes in asia: epidemiology, risk factors, and pathophysiology. JAMA, 2009;301(20):2129–2140. [PubMed] [Google Scholar]
20. Congdon N. G., Friedman D. S., Lietman T. Important causes of visual impairment in the world today. Jama, 2003;290(15):2057–2060. [PubMed] [Google Scholar]
21. Decenciere E., Zhang X., Cazuguel G., Lay B., Cochener B., Trone C., Gain P., Ordonez R., Massin P., Erginay A., et al. Feedback on a publicly distributed image database: the messidor database. 2014;33:231–234. [Google Scholar]
22. Gardner G., Keating D., Williamson T., Elliott A. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. British journal of Ophthalmology. 1996;80(11):940–944. [PMC free article] [PubMed] [Google Scholar]
23. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Elsevier. 2017 [PubMed] [Google Scholar]
24. Goh J. K. H, Cheung C. Y., Sim S. S., Tan P. C., Tan G. S. W, Wong T. Y. Retinal imaging techniques for diabetic retinopathy screening. Journal of diabetes science and technology, 2016;10(2):282–294. [PMC free article] [PubMed] [Google Scholar]
25. Graham B. Kaggle diabetic retinopathy detection competition report. 2015 [Google Scholar]
26. Gulshan V., Peng L., Coram M., Stumpe M. C., Wu D., Narayanaswamy A., Venugopalan S., Widner K., Madams T., Cuadros J., 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. [PubMed] [Google Scholar]
27. Huang L. C., Yu C., Kleinman R., Smith R., Shields R., Yi D., Lam C., Rubin D. Opening the black box: Visualization of deep neural network for detection of disease in retinal fundus photographs. The Association for Research in Vision and Ophthalmology. 2017 [Google Scholar]
28. Mookiah M., Acharya U., Chua C., Lim C., Ng E., Laude A. Computer-aided diagnosis of diabetic retinopathy: A review. In Computers in Biology and Medicine. 2013:2136–2155. [PubMed] [Google Scholar]

29. Niemeijer M., Van Ginneken B., Cree M. J., Mizutani A., Quellec G., Sanchez C. I., Zhang B., Hornero R., Lamard M., Muramatsu C., et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE transactions on medical imaging, 2010;29(1):185–195. [PubMed] [Google Scholar]
30. Oquab M., Bottou L., Laptev I., Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2014:1717–1724. [Google Scholar]
31. Wang S., Tang H. L., Hu Y., Sanei S., Saleh G. M., Peto T., et al. Localizing microaneurysms in fundus images through singular spectrum analysis. IEEE Transactions on Biomedical Engineering, 2017;64(5):990–1002. [PubMed] [Google Scholar]
32. WHO. Global report on diabetes. 2016.
33. Xu Y., Xiao T., Zhang J., Yang K., Zhang Z. Scale-invariant convolutional neural networks. arXiv preprint arXiv:1411.6369, 2014 [Google Scholar]
34. Yosinski J., Clune J., Nguyen A., Fuchs T., Lipson H. Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579, 2015 [Google Scholar].
35. 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
36. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes.JAMA. 2017;318(22): 2211-2223. doi:10.1001/jama.2017.18152
37. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology. 2018;125 (8):1264-1272. doi:10.1016/j.ophtha.2018.01.034
38. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017;124(7):962-969. doi:10.1016/j. ophtha.2017.02.008
39. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks.JAMA Ophthalmol. 2017;135(11):1170-1176. doi:10.1001/jamaophthalmol.2017.3782
40. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Med. 2018;1:39. doi:10.1038/s41746-018-0040-6
41. American Academy of Ophthalmology. International clinical diabetic retinopathy disease severity scale, detailed table. http://www.icoph. org/dynamic/attachments/resources/diabeticretinopathy-detail.pdf. Accessed 14 Oct, 2016.
42. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV. https://ieeexplore.ieee.org/document/7780677. Accessed May 12, 2019.
43. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404-413. doi:10. 1093/biomet/26.4.404
44. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy.JAMA Ophthalmol. 2013;131(3):351-357. doi:10.1001/jamaophthalmol. 2013.1743 24. Scanlon PH. Screening intervals for diabetic retinopathy and implications for care. Curr Diab Rep. 2017;17(10):96. doi:10.1007/s11892-017-0928-6
45. Wang YT, Tadarati M, Wolfson Y, Bressler SB, Bressler NM. Comparison of prevalence of diabetic macular edema based on monocular fundus photography vs optical coherence tomography. JAMA Ophthalmol. 2016;134(2):222-2
46.Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization
Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra