Literature Survey: Diabetic Retinopathy Detection

CS 574

Group 13

Group Members:

- Abhinav Hinger (160101004)
- Abhishek Suryavanshi (160101009)
- Apurva N. Saraogi (160101013)
- Daman Tekchandani (160101024)
- Divyam Agarwal (160101025)

Index

1 Abstract **3**

2 Introduction **3**

3 Approaches Explored **3**

3.1 A hybrid deep learning model for detecting diabetic retinopathy **3**

3.1.1 Introduction 3

3.1.2 Data 4

3.1.3 Methodology 4

3.1.4 Model 4

3.1.5 Results 5

3.2 The application of deep learning for diabetic retinopathy prescreening in research eye-PACS **5**

3.2.1 Data 5

3.2.2 Methodology 6

3.2.3 Preprocessing 6

3.2.4 Algorithm 6

3.2.5 Results 7

3.3 A computer-aided diagnostic system for detecting diabetic retinopathy in optical coherence tomography images **8**

3.3.1 Introduction 8

3.3.2 Methodology 9

3.3.3 Results 11

3.4 Automated Detection of Diabetic Retinopathy using Deep Learning **11**

3.4.1 Introduction 11

3.4.2 Data 12

3.4.3 Preprocessing 12

3.4.4 Methodology 12

3.4.5 Results 13

3.5 Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation 13

3.5.1 Introduction 13

3.5.2 Regression Activation Map 14

3.5.3 Dataset 16

3.5.4 Methodology 16

3.5.5 Result & Conclusion 18

4. Conclusion **19**

5. References **19**

1 Abstract

This literature survey consists of various methods to detect Diabetic RetinopathyDR)using images and automating this process. The retina images are of 2 types: 2D Fundus images and Optical Coherence Tomography(OCT) where each employ different methods to differentiate between DR and Non DR subjects.For Fundus images, different algorithms are analysed like CNN + RAM(Regression Activation Map), ResNet, CNN coupled with SVM, GoogleNet and AlexNet for classification.On the other hand, OCT images are first segmented into different retinal layers and features extracted from them which then are classified using an Autoencoder. These models are tested with different types of subject and metrics(accuracy, sensitivity, specificity etc.) are observed.

2 Introduction

The retina is the thin layer of tissue in the back of the eye that receives incoming light and sends signals to the brain through the optic nerves, these signals are what we perceive as images. Uncontrolled and chronic Diabetes can result in damage to the blood vessels in the retina. Diabetic retinopathy occurs when the damaged blood vessels leak blood and other fluids into the retina, causing swelling and blurry vision. The blood vessels can become blocked, scar tissue can develop, and retinal detachment can eventually occur. It is one of the leading causes of blindness in many countries. Skilled Ophthalmologists inspect the eye for visible signs of Retinal damage like red lesions, hard exudates, micro-aneurysms, abnormal blood vessels and haemorrhage for detection of DR. Since number of skilled ophthalmologists is less, developing an automated system which can detect DR by using retinal images will make the detection and thus the cure more accessible.

3 Approaches Explored

This section describes the papers that are analysed in terms of Introduction, Data, Methodology and Results. Various insights are derived from each usage of the models and listed below.

3.1 A hybrid deep learning model for detecting diabetic retinopathy

3.1.1 Introduction

This paper explores the use of Deep Learning for Diabetic Retinopathy detection. Writers have leveraged the automatic feature learning capability of CNNs and used it to train Linear SVM classifier.

3.1.2 Data

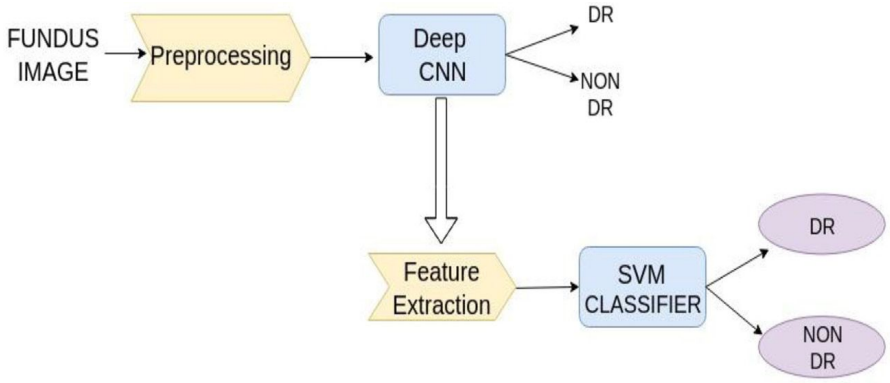
The EYEPACS dataset from kaggle is used. Since the dataset is heterogeneous (ony 9,316 out of 35,216 images are labelled as having DR) the images are mirrored and the ones having DR are also rotated by 90, 180, 270 degrees since DR detection needs to be rotationally invariant. Color balance, contrast and brightness is used for additional data augmentation.

Finally a balanced dataset with 160,386 images flattened into a 512*512 vector is used.

Non DR and DR Retina

3.1.3 Methodology

The recent success in the field of Computer Vision can be owed to the automatic feature learning capabilities of CNN. In this paper CNN is used to extract features from the images and the extracted images are then fed to a SVM classifier for training.



3.1.4 Model

The CNN used here is very similar to VGGNet.
The input size is 512*512, the number of filters double after every 2 conv layers from 32 (lower level feature extraction like lines and circles) in the first layer to 512 (higher level feature extraction like micro-aneurysms and blood vessels) in the last layer. 3*3 Kernel is used throughout the network. Convolution stride and spatial padding is fixed to 1 pixel. MaxPooling is done after every 3rd layer over a 2*2 window with stride 1.
This is followed by 3 fully connected layers having 1024 and 64 nodes and a softmax layer at the end. SGD is used with 0.4 dropout for training and regularization.
Output of the layer with 1024 nodes is used as the feature vector, this layer is the latent representation of the retinal image with most discriminative features.
The feature vectors along with labels are fed to a SVM. GridSearchCV is used for hyperparameter tuning and 7 fold cross validation is used for evaluating the model.

3.1.5 Results

The performance of the hybrid model is similar to that of standalone CNN but less images were misclassified using the hybrid model which is a very important factor considering the impact the result will have a person's life.

MODEL	SENSITIVITY	SPECIFICITY
CNN+Linear SVM	0.93	0.85
CNN	0.87	0.67

- **Sensitivity:** TP/(TP+FN)
- **Specificity:** TN/(TN+FP)

3.2 The application of deep learning for diabetic retinopathy prescreening in research eye-PACS

This paper explores the use of CNN for classification of fundus images to detect Diabetic Retinopathy.

3.2.1 Data

The public Kaggle Diabetic Retinopathy Detection competition dataset was used as the source for retinal fundus images. This dataset contains 35,126 fundus images from existing eye-
PACS users as released by the California Healthcare Foundation. The images have been labelled from 0 to 4 i.e. 5 classes in increasing order of severity.

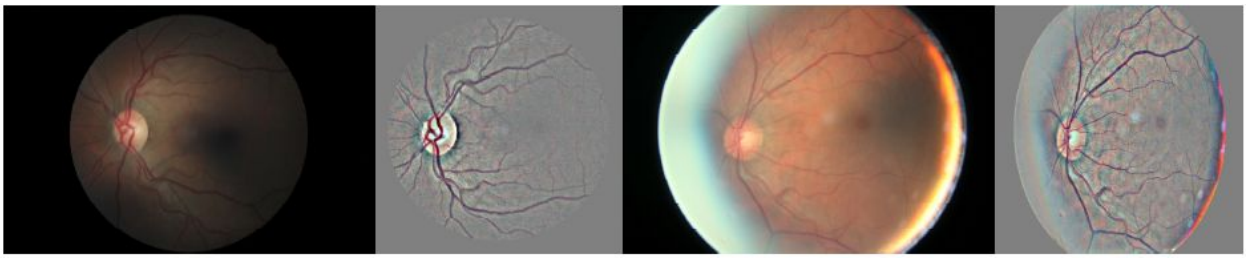
The images in the dataset have several variations in resolution and visual appearance due to use of different models and types of cameras. Some of the images are underexposed, or overexposed, out of focus and some also contain artifacts. Therefore our algorithm must address these issues effectively in order to obtain accurate results. 10% of the total 35,126 training images was set aside for our training dataset for hyperparameter selection and validation.

3.2.2 Methodology

We rescaled the images to have the same radius (~300 pixels) and then subtracted the local average color which was mapped to 50% grayscale level. Finally, we clipped the images to 90% size to remove the "boundary effects". This cropped image is then resized to 224 x 224 pixels. At runtime, the image is subtracted the VGG mean color channel intensities before feeding into the CNN20.

3.2.3 Preprocessing

The images were rescaled to have the exact radius of 300px. After that the local average color was subtracted and then mapped to 50% grayscale level. Following these steps, the images were clipped to 90% size to remove boundary effects and the final image resized to 224 x 224px.



Rating	Grade
0	No DR
1	Mild DR
2	Moderate DR
3	Severe DR
4	Proliferative DR

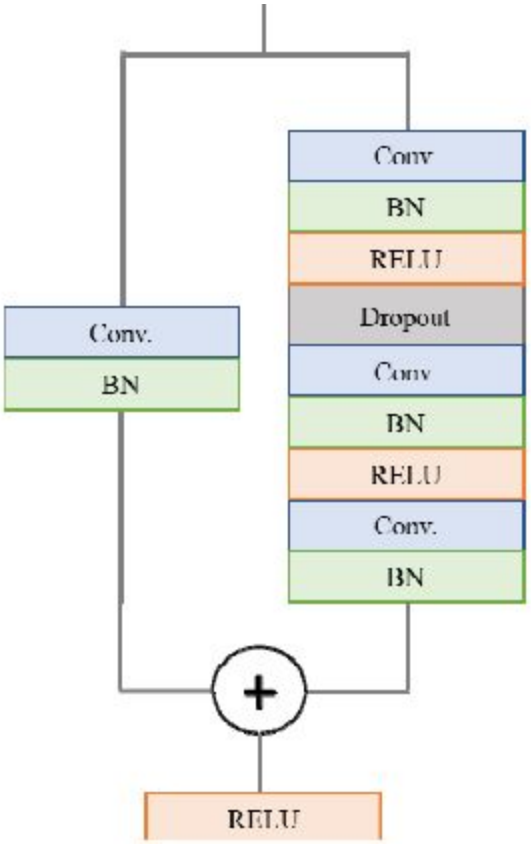
3.2.4 Algorithm

They chose Resnet-50 as the structure for the neural network.

The first convolutional layer has a kernel size of 7x7 and all other convolutions involve kernel size of 3x3 or 1x1. The initial layers try to identify primitive features such as edges whereas deeper layers try to recognize complex features such as hemorrhages and exudates. ReLU has been used as the activation function. Batch normalization has been used after each convolutional layer. A dropout layer has also been added to the convolutional block to avoid overfitting.Inputs are dropped at a rate of 50%, and dropout is added on the non-skip connection pathway of the convolutional block of ResNet-50

The weights trained for ImageNet ILSVRC algorithm were used as the initial weights for our model (transfer learning). The model was then trained with stochastic gradient descent with Adam optimization.
learning rate : 0.0001
beta-1: 0.9

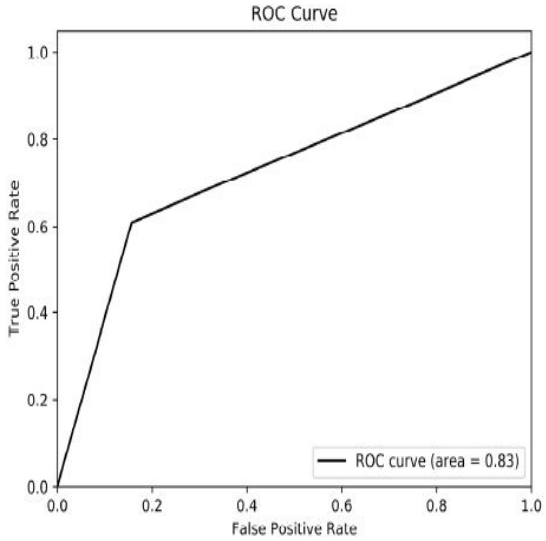
beta-2: 0.99
Epochs of training: 100(learning rate will be adjusted along with the epochs)
loss function : categorical cross-entropy function

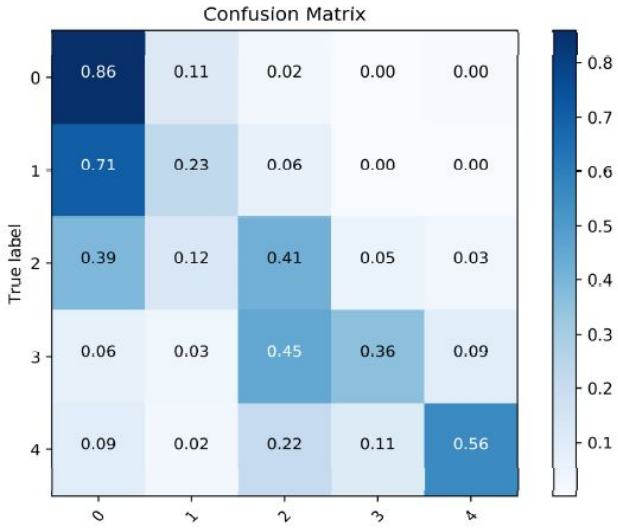


3.2.5 Results

Quadratic weighted Kappa(Kaggle metric) of 0.64 was achieved by our model. AuROC (area under receiver operating curve) of 0.83 was achieved. Also specificity of 84% and 61% sensitivity was achieved across the full dataset. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of .5 represents a worthless test

- *Sensitivity*: probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage).
- *Specificity*: probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage).



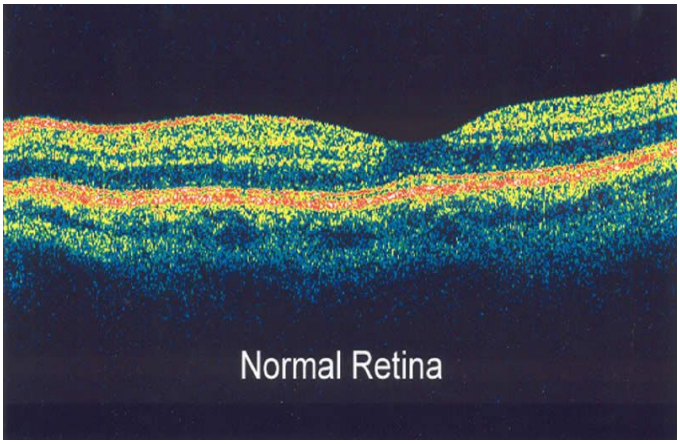


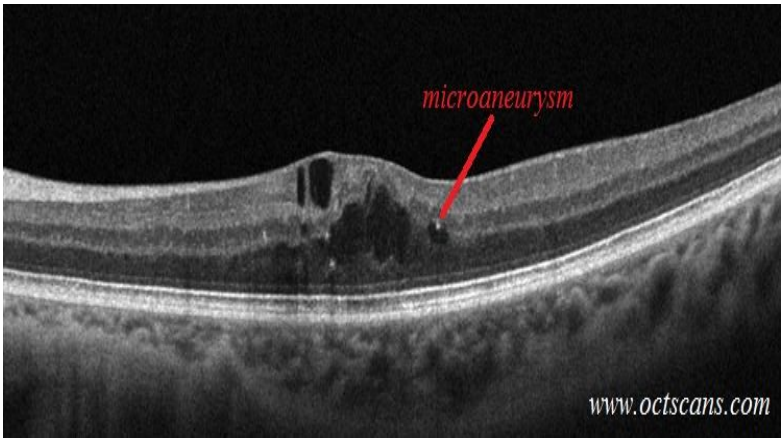
True Positive rate = sensitivity
False Positive rate = 1 - specificity

3.3 A computer-aided diagnostic system for detecting diabetic retinopathy in optical coherence tomography images

3.3.1 Introduction

Optical Coherence Tomography Images:
OCT is an imaging technique that uses light waves to take cross sectional images of the retina. Using this each distinct layers of retina can be observed by the ophthalmologist and the thickness of each layer can be measured. This process of segmenting the layers is difficult to automate and is even more difficult for where the image is blurry or in other words the image has low SNR(Signal to Noise ratio).





This paper does the automatic diagnosis in three steps:

1. Segmenting and Localizing an OCT image into 12 layers.
2. Individual layer is characterized by a 3 features: Reflexivity, Curvature and Thickness in terms of their Cumulative Probability distributions.
3. A pre trained Deep fusion classification network(DFCN) finds the most discriminating features.

Since Computer vision based DR detection of fundus images works well for cases of No DR and higher severity DR but fails to detect early stage DR because the size of lesions is small. This paper detects early stage DR with good accuracy, sensitivity and specificity. Another goal was to decrease the segmenting errors in OCT images. 2D Fundus images have the drawback of lacking any depth information of the retina and are not cost effective. Decreasing retinal thickness and optical reflectivity are significant biomarkers for detecting DR changes with OCT

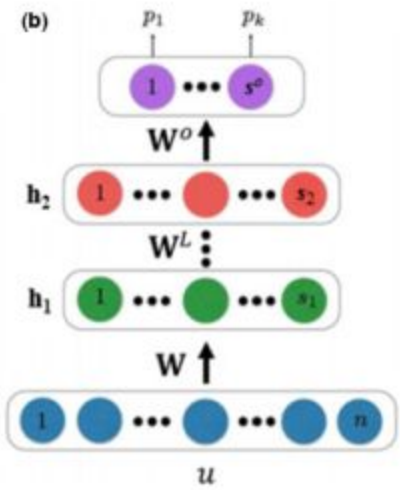
3.3.2 Methodology

An input OCT image, **g** first is aligned to a training database using some key points like the fovea, and its region map, **m** are described with a joint probability model. This conditional probability is defined using an unconditional probability which in turn uses an Adaptive Shape Prior.
This is created using a **Shape Prior** database of OCT scans of 12 humans(6M,6F). The shape prior can be used to segment the aligned image using the control points.

Pixel wise intensities of the input grayscale image is normalised to get the Empirical marginal probability distribution. This is then approximated by adaptive linear combination of sign-alternate discrete Gaussians (**LCDG**) which is the current state of the art in biomedical **segmentation**.

Above obtained shape prior and intensity model is combined with Markov Gibbs Random Field (MGRF) of spatial interaction of pixel wise labels. This is done to better account for noise in data and homogeneity of the data.

After we obtain the segmented OCT scans, the features: **reflectivity**, **curvature** and **thickness** are extracted. For each subject, these three features of the extracted retina layers are described as a whole with a cumulative distribution function (CDF). This CDF globally identifies the discriminatory features.
These CDFs are used as input to the classifier which is a stack of non negatively constrained auto Encoders (**SNCAE**) followed by output softmax(Probability of DR and Non DR) as given below.
Note that we cannot use the conventional classification techniques because of different sizes for each subject and also taking large amount of time to train. The network fuses the CDFs into their **latent description** which is lower dimensional and learns to classify OCT images.



This is an example of SNCAE with 2 NCAE layers and an output softmax layer. Encoding layer converts the **n**-dimensional vector into a lower dimensional vector **h** and Decoding layer converts it back to n-dimensional vector. Weights and activations are learned so that it can effectively convert a lower dimensional vector back into the original signal.

To decrease the number of negative weights and enforce sparsity of coding, a non negativity constrained AE (**NCAE**) is built by adding penalty for negative weights and non-zero codes.

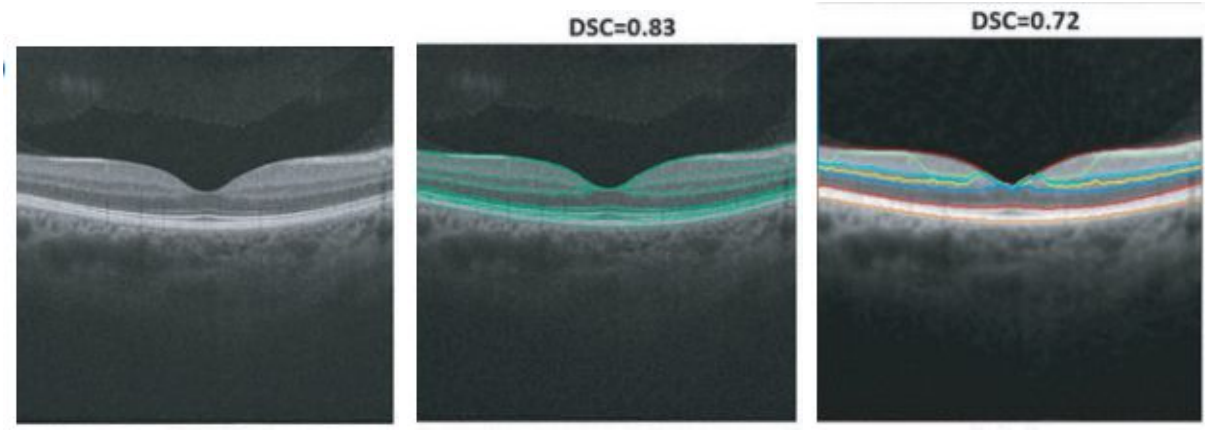
$$J_{\mathrm{NCAE}}(\mathbf{W}) = J_{\mathrm{AE}}(\mathbf{W}) + a\sum_{j=1}^{s}\sum_{i=1}^{n} f(w_{j:i}) + bJ_{\mathrm{KL}}(\mathbf{h}_{\mathbf{W}};\gamma)$$

The Architecture is built by adding 2 NCAE layers which are already pre trained in an unsupervised manner. It is further fine tuned by taking penalty for only negative weights.
The pretraining and fine tuning of each layer individually gives as probability output top most codes $\mathbf{h}_f$.

$$p_t(c;\mathbf{W}^t_{o:c}) = \frac{\exp\left((\mathbf{W}^t_{o:c})^\top \mathbf{h}_t^{[L]}\right)}{\exp\left((\mathbf{W}^t_{o:1})^\top \mathbf{h}_t^{[L]}\right) + \exp\left((\mathbf{W}^t_{o:c})^\top \mathbf{h}_t^{[L]}\right)}$$

3.3.3 Results

This Computer Aided Diagnostic system was tested on 52 people: 26 normal and 26 abnormal.
The robustness and accuracy of the approach is measured using Agreement Coefficient(AC) and Dice Similarity Coefficient(DSC). Higher DSC scores were obtained than previous segmentation approaches on both images with high and low SNR.



The classifier decides whether the DR or no DR, based on the CDFs of the most discriminative features (the INL curvature, MZ reflectivity, and NFL thickness).
The classifier was tested with **leave-one-out cross-validation** test with all the 52 OCT images where it achieved 100 percent accuracy.

Classifier	Accuracy	Sensitivity	Specificity	AUC
DFCN (proposed)	**100%**	**100%**	**100%**	**0.98**
K-Star (K*)	95%	95%	95%	0.94
K-Nearest-Neighbor (K-NN)	90%	90%	90%	0.90
Random forest	85%	85%	85%	0.87
Random tree	72%	72%	72%	0.71

Classifier	Training accuracy (fourfold cross-validation)	Testing		
		Accuracy	Sensitivity	Specificity
DFCN (proposed)	**95%**	**92%**	**83%**	**100%**
K-Star (K*)	93%	89%	89%	89%
K-Nearest-Neighbor (K-NN)	91%	84%	84%	83%
Random forest	85%	82%	82%	82%
Random tree	83%	81%	81%	81%

Another experiment included choosing 40-12 training-test split and doing a **4 fold cross validation**. In both tests, the DFCN outperformed conventional classification methods like k-Nearest Neighbours, Random forest etc.

3.4 Automated Detection of Diabetic Retinopathy using Deep Learning

3.4.1 Introduction

This paper explores the use of Convolutional Neural Network(CNNs) to detect Diabetic Retinopathy particularly focussing on Early stage retinopathy. Detecting Early stage DR is very difficult since the features which signify it are very subtle and small in size like microaneurysms which are not detected by basic models. Most errors occur

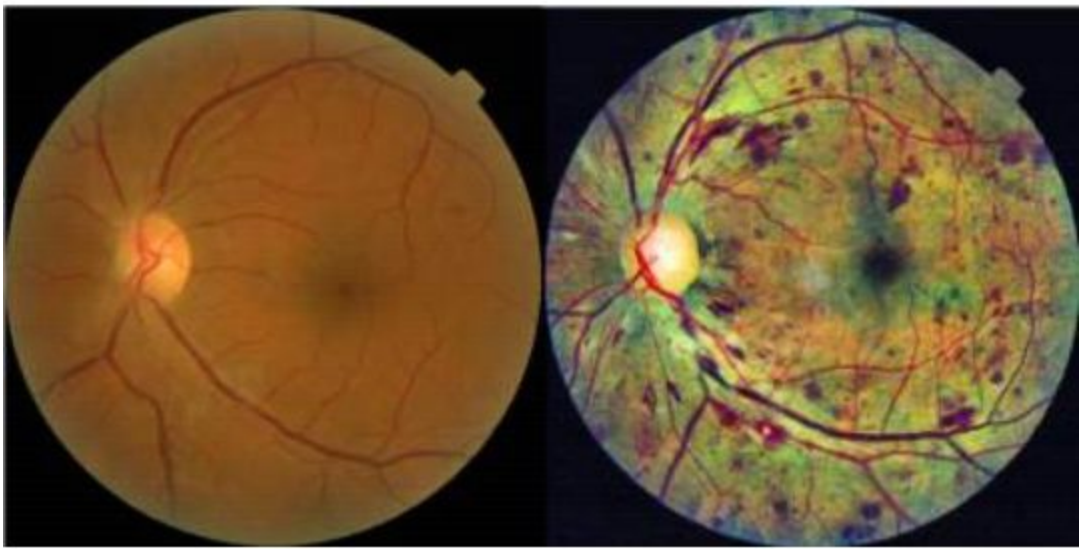
when classifying mild disease as normal. This paper used transfer learning on models GoogLeNet and AlexNet which were trained on ImageNet dataset to do a 2-nary, 3-nary and 4-nary classification and measured the metrics. This paper seeks to increase the sensitivities of mild/early DR.

3.4.2 Data

This paper used the **Kaggle** dataset of 35,000 retinal images with 5-class labels (normal, mild, moderate, severe, end stage) and a physician-verified **Messidor-1** dataset of 1,200 color fundus images with 4-class labels. The Kaggle images had much higher noise and Variance in resolution, lighting conditions, colour etc. In the interest of efficient model building, progressed to a smaller but more ideal dataset for learning difficult features. The Messidor dataset was supplemented with a Kaggle partition (MildDR) consisting of 550 images which were of higher quality and were verified by ophthalmologists.

3.4.3 Preprocessing

Since there are many images with too much black area which is uninformative and hinders the learning of the convolutional network, these are removed using Otsu's method. Images are then normalised and Contrast Limited Adaptive Histogram Equalisation(CLAHE) is applied to get better discernible features.



Before and After preprocessing

3.4.4 Methodology

GoogLeNet is 22 layer deep, very efficient network that achieves state-of-the-art accuracy using a mixture of low-dimensional embeddings and different sized spatial filters. Better use of internal network and increased depth of convolutional layers allows the model to learn deeper features. BatchNorm and Dropout layer after every Conv layer is used and MaxPooling is done using kernel size 3x3 and stride 2. LeakyRelu is used as activation function, L2 Regularisation and Cross Entropy loss to decrease Overfitting.
Also Weights are not initialised randomly but using Xavier Method which increases the chances of convergence and quickly. Images are used 256x256 as Input to the models previously trained on ImageNet database

3.4.5 Results

CLAHE technique applied using OpenCV as a preprocessing step increasing the accuracy considerably for all types of classification. For eg. 3-ary classifier sensitivity for the mild case increased from 0 to 29.4%

2-nary classification:

GoogLeNet model got sensitivity of 95% and specificity of 96% using data augmentation and preprocessing techniques and matched the SOTA in the field.

3-nary and 4-nary classification:

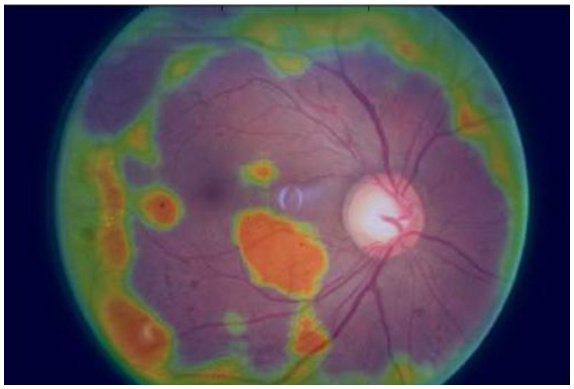
Model performed well for no DR and severe DR but achieved sensitivity of only 7%. It was observed that performance is limited by not detecting subtle features by CNN

Visualizing the Results as Heat Map

Using heat maps to mark high level and low level features which are detected by CNN gives us insights like bigger size features and different colors are easily detected by CNN.

4-nary classifier:

It was observed that there was just not enough data for GoogleNet to classify the images. It was underfitting and acted as a Majority classifier.



Thus, we can observe CNN can be leveraged with large datasets for disease screening. High Bias and low variance means CNN can be applied to other diseases as well. However, the only features which CNN detected were which were easily observable and ignoring the more subtle features. Mild vs Normal classification involved features which were less than 1% of the pixel volume.

3.5 Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation

3.5.1 Introduction

They proposed a deep learning method for interpretable diabetic retinopathy (DR) detection. By adding the regression activation map (RAM) after the global averaging pooling layer of the convolutional networks (CNN) we can highlight the regions of interest through which we actually identified the severity of the disease.

Dataset which they used consists of retina images captured using color fundus photography. Feature extraction on this kind of photography is more challenging than the normal photography methods we are used to as the key signals are generally tiny and indiscriminating from the noise in fundus photography. Therefore it is important to develop a systematic feature detection method to characterize the type of features particularly related to diabetic retinopathy detection task.

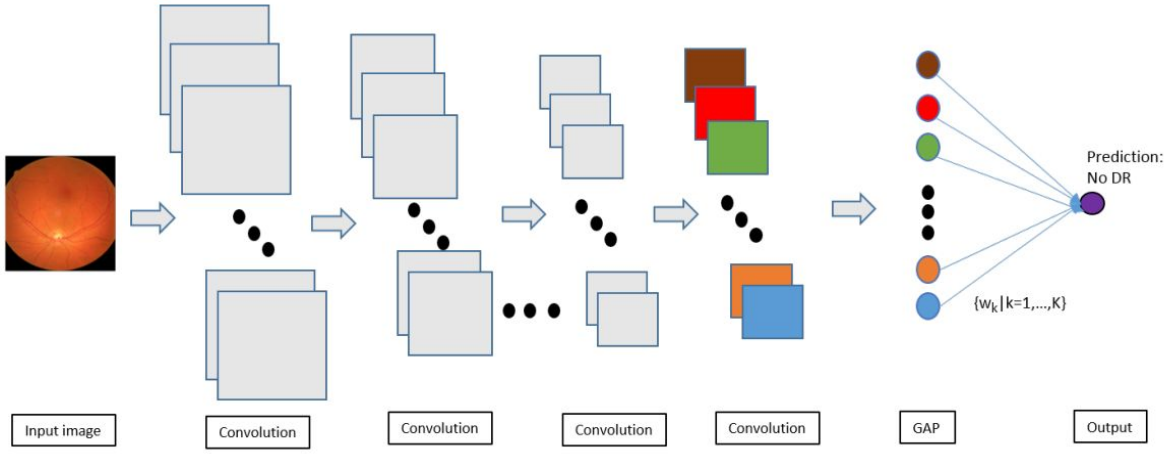
In this paper they used CNN which has no fully connected layer and only have convolutional and pooling layers. This reduces the number of parameters significantly as fully connected layers generally brings more parameters than convolutional layers in conventional CNN which helps us in interpreting the neural network better. Also in the experiments they have shown that prediction performance of their network is comparable to the one having fully connected layers. Major advantage of using the network proposed by them is that it can provide RAMs of input image to show the contribution score of each pixel of the input image for DR detection task.

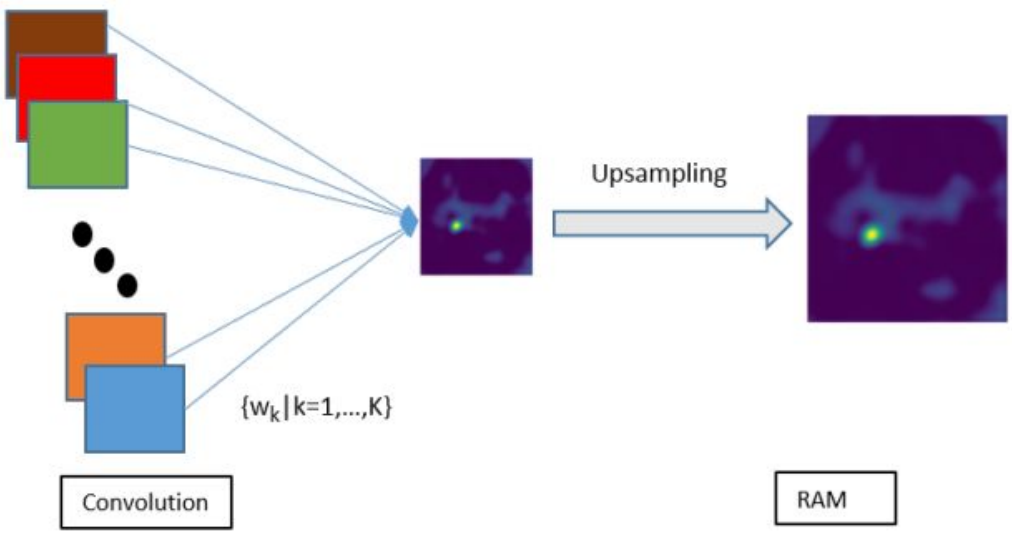
Previously generic feature extraction methods like hough transform, gabor filters and intensity variations, etc. were used for feature extraction. Object detection algorithms like support vector machines and k-NN were used with these extracted features to identify and localize exudates and hemorrhages.But these kind of approaches are not as effective as deep learning approaches in terms of prediction performance. Also the DR competition on kaggle saw almost every top submission using CNN as their key algorithm.

3.5.2 Regression Activation Map

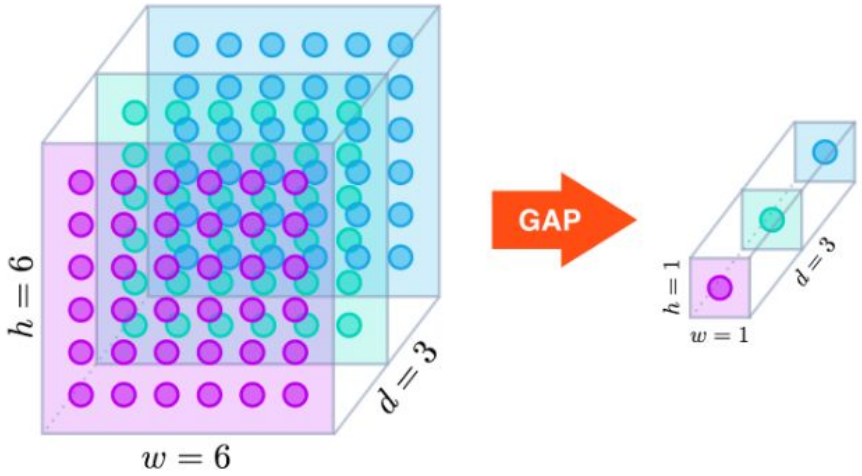
In this paper they propose to generate RAM of an input image to localize the discriminative regions or point out the regions of interest towards the regression outcomes.

Here they have used GAP and linear output unit to visualize the region of interest according to given regression value. They are not using fully connected layers as it they make it difficult to identify the importance of different units for identifying the output labels. Proposed network used Global Average Pooling to connect last convolutional layer with output layer.





GAP layers perform an extreme type of dimensionality reduction, where a tensor with dimensions h×w×d is reduced in size to have dimensions 1×1×d. GAP layers reduce each h×w feature map to a single number by simply taking the average of all hw values.



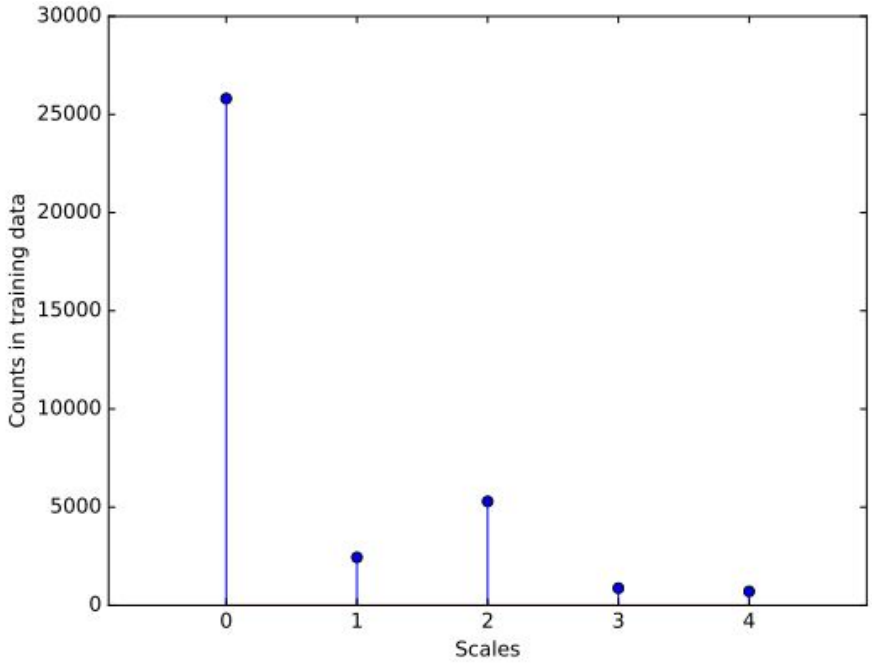
Activation maps in the final convolutional layer before GAP layer acts as a detector of different local pattern in the image. Transforming these detected patterns to detect objects we can get class activation map for an image.

This transformation is done by noticing each node in the GAP layer corresponds to a different activation map, and that the weights connecting the GAP layer to the final dense layer encode each activation map's contribution to the predicted object class. To obtain the class activation map, we sum the contributions of each of the detected

patterns in the activation maps, where detected patterns that are more important to the predicted object class are given more weight.

3.5.3 Dataset

The dataset (link) used is from the kaggle competition on predicting DR. Dataset contains 35126 high resolution images under a variety of imaging conditions. For every subject we have images of his both left and right eye. The labels were provided by clinicians who rated the presence of diabetic retinopathy in each image by a scale of “0, 1, 2, 3, 4”, which represent “no DR”, “mild", “moderate”, “severe”, “proliferative DR” respectively.



Data Imbalance between DR and Non DR

3.5.4 Methodology

First, they sampled all classes such that all classes are represented equally on average.

Initialization and pretraining Orthogonal initialization is used to initialize weights and biases. Using 128 pixel images they first trained a smaller network. Initializing the medium network from the previously trained weights from smaller networks. This medium network used 256 pixel images. Finally the large network was trained on 512 pixel images initialized with trained weights of medium network.

The common image transformations like translation, rotation, flipping, stretching and color augmentation are used for data augmentation.

Leaky rectifier units (0.01) following each convolutional layer were used for non linearity. Using Nesterov momentum networks were trained with fixed schedule over 250 epochs. For the nets on 256 and 128 pixel images, we stop training after 200 epochs.

L2 weight decay with factor 0.0005 are applied to all layers. Loss function is mean squared error, as we treat this problem as a regression one. The convolutional networks have untied biases. Batch size is fixed at 32 for all networks. Thresholds to discretize regression values to obtain integer levels for computing Kappa scores.

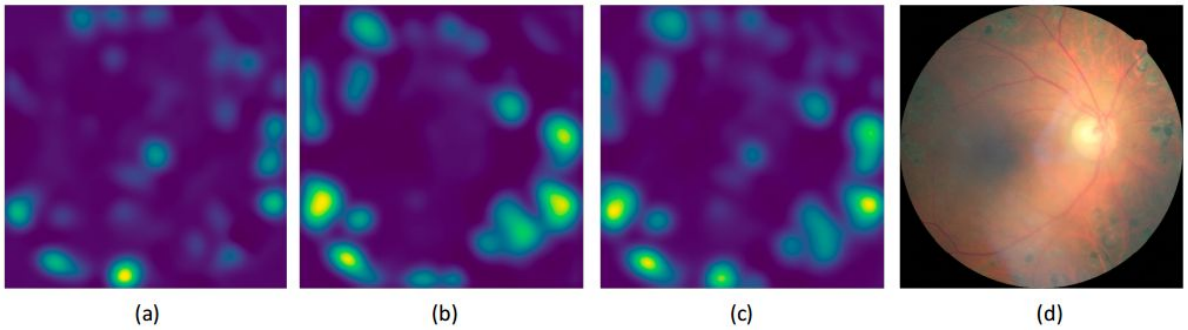
Severity of Disease	Thresholds (Predicted regression values)
No DR	< 0.5
Mild	>=0.5 and <1.5
Moderate	>=1.5 and <2.5
Severe	>=2.5 and <3.5
Proliferative	>=3.5

Instead of using fully connected layer they used GAP layer which reduces their parameter size by 21.8% and speed-up the training by 11.8%-13.1% and still achieving comparative Kappa score to the benchmark method.

Performance statistics of the benchmark and our approaches on test dataset.

	Baseline	Ours
Kappa score (Public Leaderboard)	0.8542	0.85034
Kappa score (Private Leaderboard)	0.8448	0.8412
Parameter # (net-5)	12.4M	9.7M
Training time (second/epoch)	422.1	367.3
Parameter # (net-4)	12.5M	9.8M
Training time (second/epoch)	451.7	398.2
RAM	No	Yes

They used Net-5 for generation of RAM with 128 and 256 pixel images as input. To improve localization ability of RAM several layers of Net-5 were removed for different input sizes as it will increase spatial resolution of the last convolutional layer before GAP.



RAMs were generated on both 128 and 256 pixel images as shown in fig (a) and (b) respectively and their average is shown in ©. We can argue that (c) captures the ROI in better manner as compared to the other two if refer to the original image (d). Therefore they fused together RAMs generated on different sizes of images to get better ROI and only fused RAMs were reported in the study.

3.5.5 Result & Conclusion

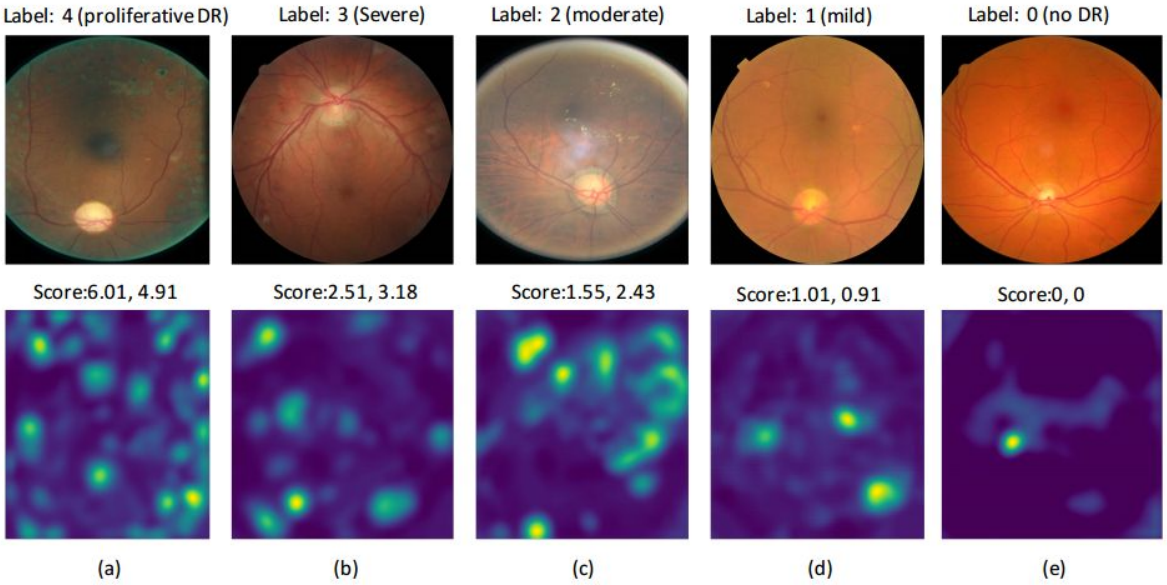


Figure 5. Ground truth and the corresponding RAMs. The two scores are from the 128 and 256 pixel images, respectively.

RAM provides the reasonable transparency on our deep learning model to see why and how it makes the decision. The visual explanation of RAM may assist clinicians to quickly identify the pathogenesis of disease. RAM layer can provide the robust interpretability of the proposed detection model by monitoring the pathogenesis so that the proposed model can be taken as an assistant for clinicians.

4. Conclusion

We analysed 2 types of data 2D fundus and OCT scans. DIfferent methods were analysed for both datasets. Most of the CNN based methods accurately differentiated between Proliferative DR and No DR. Also CNN based methods identified features which were observable but could not identify subtle features. Also localisation of features was achieved using CNN with RAM. Early stage DR can be identified using OCT scans by segmenting the Retinal layers and identifying discriminating features which gave satisfactory performance.

5. References

5.1 A hybrid deep learning model for detecting diabetic retinopathy - https://www.tandfonline.com/doi/abs/10.1080/09720510.2018.1466965

5.2 The application of deep learning for diabetic retinopathy prescreening in research eye-PACS - Siliang Zhang, Huiqun Wu, Veda Murthy, Ximing Wang, Lin Cao - https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10579/1057913/The-application-of-deep-learning-for-diabetic-retinopathy-prescreening-in/10.1117/12.2296673.full?SSO=1

5.3 A computer-aided diagnostic system for detecting diabetic retinopathy in optical coherence tomography images - ElTanboly A, Ismail M, Shalaby A, Switala , El-Baz, Schaal S, Gimel'farb G, El-Azab - https://www.ncbi.nlm.nih.gov/pubmed/28035657

5.4 Automated Detection of Diabetic Retinopathy using Deep Learning - Carson Lam, Darvin Yi, Margaret Guo, Tony Lindsey - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961805/

5.5 Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation - Zhiguang Wang, Jianbo Yang - https://arxiv.org/abs/1703.10757