Diabetic Retinopathy Classification of Fundus Images

A Project Report

Submitted for Minor Project (7CS193) of 7th Semester for partial fulfilment of the requirements for the award of the degree of

Bachelors of Technology in
Computer science and Engineering

Submitted by

Aditya Ankul 1706020
Anurag Kumar 1706023
Piyush Manohar 1706027
Guddoo Prasad 1706029

Under the Supervision of

Dr A. S. Tewari
Asst. Prof. CSE Department
NIT Patna

Department of Computer Science & Engineering

National Institute of Technology Patna
Ashok Rajpath, Patna-800005.

Sep-Nov, 2020

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

CERTIFICATE

This is to certify that Aditya Ankul Roll No. 1706020, Anurag Kumar Roll No. 1706023, Piyush Manohar Roll No. 1706027, Guddoo Prasad Roll No. 1706029 has carried out the Minor project (7CS193) entitled as "Diabetic Retinopathy Classification for Fundus Image" during their 7th semester under the supervision of Dr A. S. Tewari, Asst. Prof., CSE Department in partial fulfilment of the requirements for the award of Bachelor of Technology degree in the department of Computer Science & Engineering, National Institute of Technology Patna.

…….
Dr A. S. Tewari
Assistant Professor
CSE Department
NIT Patna

…….
Dr J. P. Singh
Head of Department
CSE Department
NIT Patna

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

DECLARATION

We students of 7th semester hereby declare that this project entitled "Diabetic Retinopathy Classification of Fundus Image" has been carried out by us in the Department of Computer Science and Engineering of National Institute of Technology Patna under the guidance of Dr A. S. Tewari, Department of Computer Science and Engineering, NIT Patna. No part of this project has been submitted for the award of degree or diploma to any other Institute.

Name	Signature
Aditya Ankul	…….
Anurag Kumar	…….
Piyush Manohar	…….
Guddoo Prasad	…….

Place:
NIT Patna

Date:
……..

ACKNOWLEDGEMENT

We would like to acknowledge and express my deepest gratitude to our mentor Dr A. S. Tewari, Assistant Professor, Computer Science & Engineering Department, National Institute of Technology Patna for the valuable guidance, sympathy and co-operation for providing necessary facilities and sources during the entire period of this project.

I wish to convey my sincere gratitude to the Head of Department and all the faculties of Computer Science & Engineering Department who have enlightened me during our studies. The faculties and cooperation received from the technical staff of Department of Computer Science & Engineering is thankfully acknowledged.

1. Aditya Ankul (Roll No. 1706020)
2. Anurag Kumar (Roll No. 1706023)
3. Piyush Manohar (Roll No. 1706027)
4. Guddoo Prasad (Roll No. 1706029)

Contents

1. Abstract:

Retina is the outer lining of human eye where the image formation takes place. Any threat to retina causes severe eye defects and may lead to complete blindness. During a defect the retina gets distorted and several other complications may arise.
To measure the severity of a disease we need to determine different retinal tissue damages. These damages must be quantified to make useful predictions. Here we attempt to quantify retinal tissue damage through various image processing techniques.
To verify our estimate, we applied machine learning algorithms to create a classifier for detection of diabetic retinopathy and macular edema disease.
Diabetic retinopathy & Macular Edema are diseases prone to diabetic people. They cause progressive damage to the retina of the eye. DR is the leading cause of blindness in the working-age population of the developed world. It is estimated to affect over 93 million people.

2. About the Project:

2.1 Introduction:

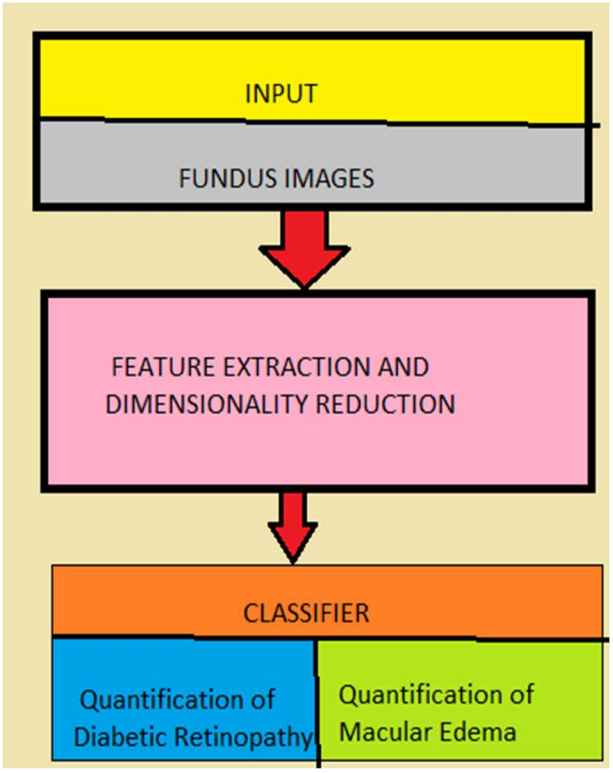
To study a retina examination is done. The images of retina are taken through fundus photography. Fundus photography involves capturing a photograph of the back of the eye. Fundus images are then analyzed by ophthalmologists who look for certain patterns and defects in the image to predict diseases.

There are many problems in this system. World is short of highly qualified ophthalmologists. Due to this people have to wait for long before starting medications. This sometimes worsens the condition. Another crucial disadvantage is lack of agreement between different doctors on a single profile of fundus.

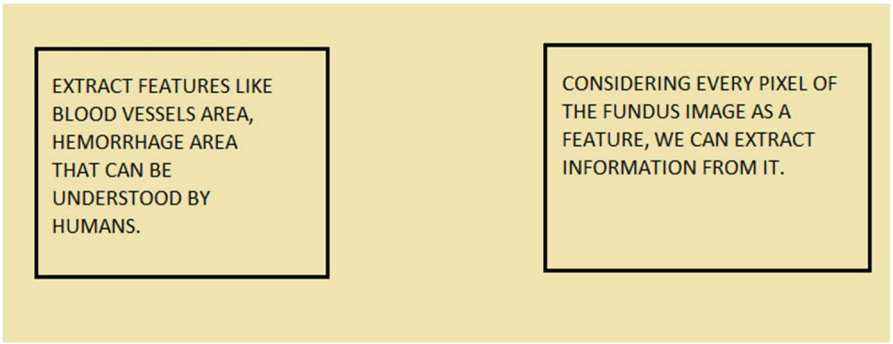
Our project aims to analyze these defects through sophisticated image processing techniques. Based on known patterns and defects we extract features from fundus image. These features are taken and put into a classifier. The classifier comes out with a decision based on its learning.

2.2 Methodology

Methodology of the project can be overseen as follows:



The feature extraction can be done by two methods:



3. Implementation

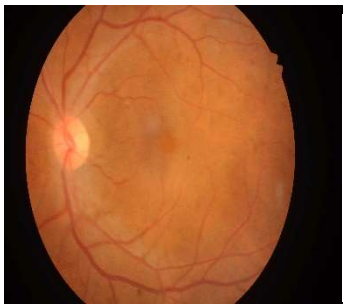
Method 1: By extracting features understandable by humans

Clinicians can identify DR by the presence of lesions associated with the vascular abnormalities caused by the disease. While this approach is effective, its resource demands are high. The expertise and equipment required are often lacking in areas where the rate of diabetes in local populations is high and DR detection is most needed. As the number of individuals with diabetes continues to grow, the infrastructure needed to prevent blindness due to DR will become even more insufficient.

The need for a comprehensive and automated method of DR screening has long been recognized, and previous efforts have made good progress using image classification, pattern recognition, and machine learning.

Features of Fundus Image that were extracted:

- **Detection of Blood Vessels:** Blood vessels are very important features which help in examination of retina. Blood vessels show different symptoms when an eye suffers from certain disease. There are swellings in the blood vessel when an eye has diabetic retinopathy. So, detection of blood vessels becomes very important. We have used Alternate sequential Filtering (ASF) along with other image processing techniques to extract blood vessels. We also tried to work on red channel of image and get it segmented. But the presence of hemorrhages and clots made it difficult. In our procedure we have extracted green channel of image because it has greater contrast. To further increase contrast, we apply Contrast Limited Adaptive Histogram Equalization. Applying ASF on this image gives us another image with average intensity of each region applied over it. Later we subtract this image from output of CLAHE. This gives us an image which contains faint traces of blood vessels with optic disk and other things removed. We binarize this image with a threshold T and get blood vessels segmented. The final image also contains noise and some undesirable elements. Noise is removed by eroding the image. Undesirable elements are removed by taking into account the feature that only blood vessels are linear in shape.



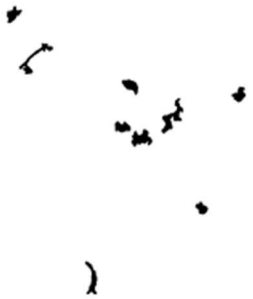
Input image of fundus

Segmented blood vessels

- **Detection of Hemorrhages:** Hemorrhages are chunks of blood vessels lying on the retina because of leakage. Presence of Hemorrhages are a good indicator of retinal damage. In a fundus image hemorrhages appear as red blobs. To detect hemorrhages, we followed the same procedure as

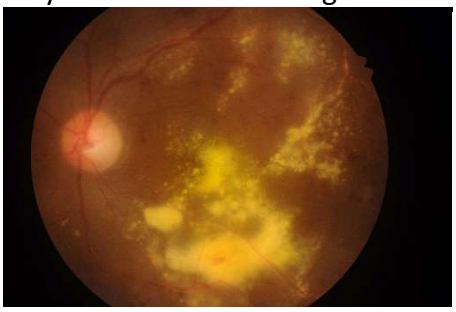
we did in blood vessel detection up to some steps. When we get the image (f) we do XOR operation on current image and image obtained after segmenting blood vessels. This gives us an image which contains hemorrhages and small noise elements. To remove noise, we apply a median filter on image. The new image has hemorrhages.

Fundus image



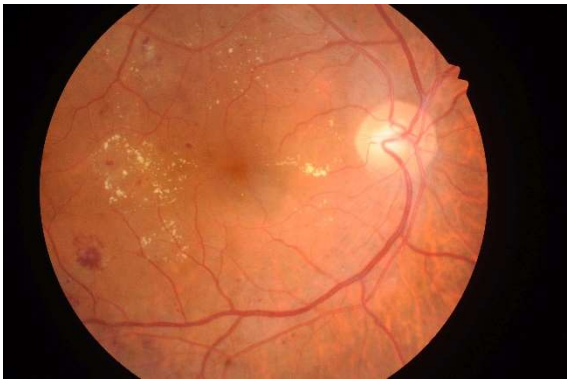
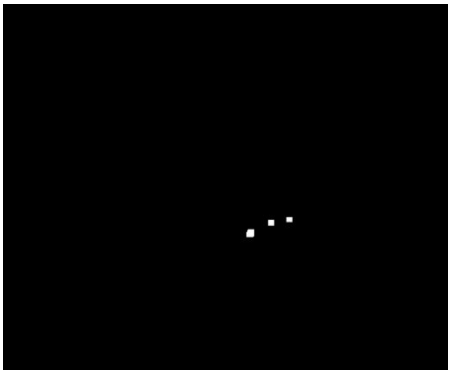
its hemorrhages

- **Detection of Exudates:** Exudates are another important feature of retinal analysis. Exudates are bulges of yellow and white colors appearing in the fundus. Exudates have been detected using red and blue channel. We know that yellow color consists majorly of red and green color. We threshold the red channel and green channel individually taking threshold T as (max(channel)+mean(channel))/2. These two binary channels were passed through AND gate which gives us all the yellow-colored pixels. Now in the image we have exudates along with optic disk. We make a window of suitable size which scans the image with a stride s. For every scan we calculate the mean intensity of every window. Since optic disk is bigger in size and has very high intensity, we get the maximum at optic disk. To remove it we mask it with a window of black color. Thus, we get only exudates in our image.



Fundus and its segmented exudates

- **Detection of Microaneurysm:** Microaneurysm occur as small dark round dots (~ 15 - 60 mm) on fundus images. They are small bulges developed on weak blood vessels and are earliest sign of DR. The green channel of the image is extracted as it gives the best contrast between the microaneurysms and other bright parts such as optic disk, exudates, etc. Now, image contrast is further stretched by applying adaptive histogram equalization (CLAHE). The image contrast is stretched by applying adaptive histogram equalization before using edge detection (Canny method) to detect the outlines of the image. The edge detection image is then subtracted from the image with boundaries to obtain an image without boundaries. After that, the holes or gaps are filled, resulting in microaneurysms.

Fundus and its segmented microaneurysm

- **Calculation of Entropy:** Entropy is a statistical measure of randomness that can be used to characterize the texture of the input images. If image has more than two dimensions, the entropy function treats it as a multidimensional grayscale image and not as an RGB image. Entropy is defined as:

$$E = -\sum_{i}\sum_{j}(p * log_2 p)$$

where p is histogram values of gray scale image at different (i, j).

- **Calculation of Homogeneity:** Homogeneity is another feature that is used to examine the texture of the retinal image which is calculated using the Gray-Level Co-Occurrence Matrix (GLCM). GLCM is a statistical method of examining texture that considers the spatial relationship of pixels. It is created by calculating how often pairs of pixels with specific values and in a specified spatial relationship occur in an image. Homogeneity weights values by the inverse of the contrast weight, with weights decreasing exponentially away from the diagonal as shown in the following equation. The addition of value '1' in the denominator is to prevent the value '0' during division. As homogeneity increases, the contrast typically decreases

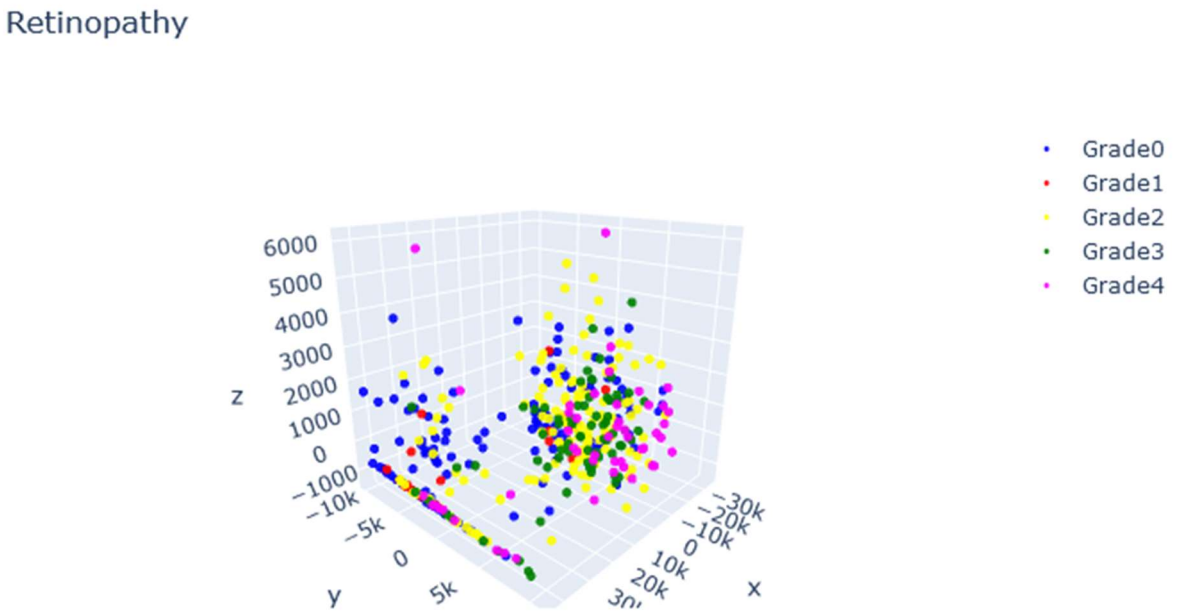
$$H = \sum_{i}\sum_{j}\frac{1}{1+(i-j)^2}p_d(i.j)$$

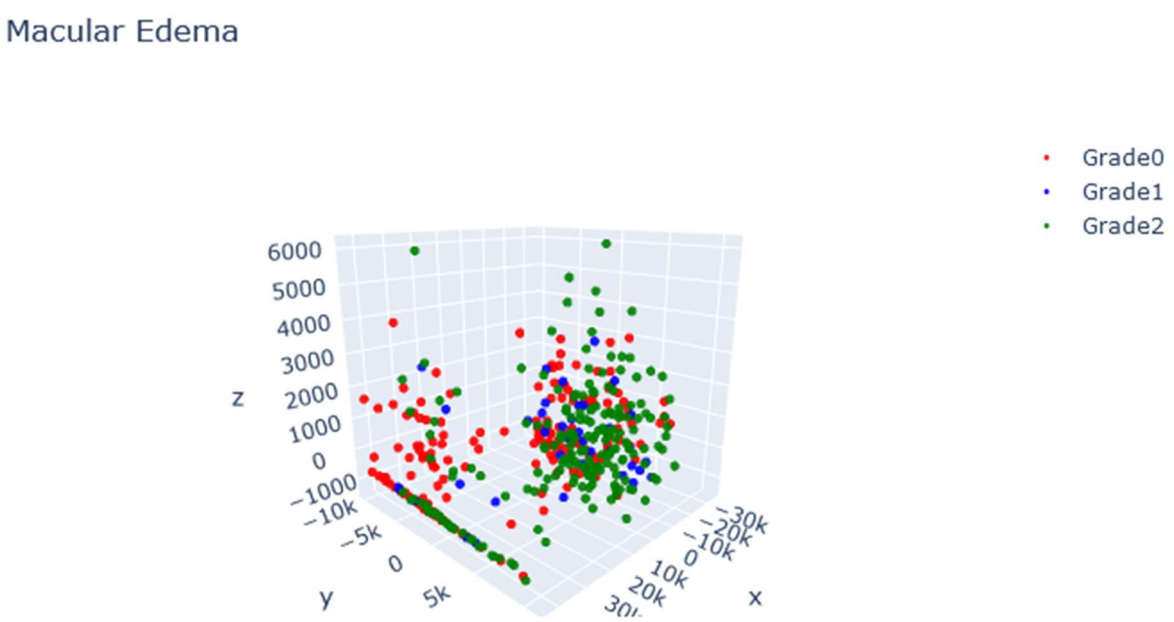
Data:

Our data comprises of numerical values of these 12 features. Converting images into numerical data is the first step in machine learning analysis. Now, the converted data was cleaned by l1 normalization and missing/ambiguous values were filled using median or mean for specific features like FD, etc.

Image nan	Retinopath	Risk of ma	ma	bva	ha	hmg	ertp	SD	HUE	SATU	INTN	RED	GREEN	HUE_M
IDRiD_001	3	2	1506	7178	3312	0.322421	10.36915	0.441901	0.178452	0.721233	0.360476	0.472828	0.305247	-2.55E-05
IDRiD_002	3	2	1917	6074	5475	0.3327	10.07208	0.440776	0.176194	0.735644	0.351551	0.466491	0.290952	-2.42E-05
IDRiD_003	2	2	274	6592	32047	0.407565	9.392455	0.408374	0.147216	0.865893	0.307236	0.441078	0.228001	7.48E-05
IDRiD_004	3	2	1696	6122	4555	0.344969	10.44417	0.427901	0.240384	0.538559	0.327029	0.390426	0.255746	5.24E-06
IDRiD_005	4	0	2526	7429	3145	0.340237	10.32779	0.442558	0.186319	0.478911	0.391841	0.439139	0.368257	-4.49E-08
IDRiD_006	4	1	754	9838	11573	0.363883	10.13358	0.429688	0.167804	0.661687	0.316288	0.393374	0.230239	7.00E-06
IDRiD_007	4	0	2013	543	5168	0.364361	10.37336	0.430947	0.183376	0.65949	0.324663	0.422936	0.252826	-4.04E-06
IDRiD_008	4	2	2306	3781	4659	0.364054	10.09071	0.413562	0.187908	0.610309	0.30908	0.349645	0.239431	1.95E-05
IDRiD_009	3	2	519	4365	12185	0.363781	10.04762	0.410699	0.176596	0.674098	0.284496	0.333254	0.206467	-8.11E-06
IDRiD_010	4	1	502	6191	9910	0.368139	9.790815	0.409455	0.164446	0.84718	0.316792	0.483333	0.218144	3.38E-05
IDRiD_011	3	1	1596	2073	24161	0.356301	9.720703	0.41585	0.160522	0.807761	0.314062	0.429636	0.243212	6.41E-05
IDRiD_012	3	2	1412	7925	303	0.408566	9.479227	0.425165	0.154423	0.583975	0.396521	0.534981	0.329175	-1.68E-05
IDRiD_013	3	0	441	7980	17	0.42694	9.578874	0.410642	0.146485	0.659888	0.379934	0.529321	0.304381	-1.16E-05
IDRiD_014	4	2	969	4125	34676	0.368283	9.847142	0.421432	0.158611	0.840935	0.271672	0.329386	0.191333	0.000133

Table: Dataset made using features of fundus image.





Principal Component Analysis: The main goal of a PCA analysis is to identify patterns in data; PCA aims to detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense. In a nutshell, this is what PCA is all about: Finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information. PCA generates at most n number of features corresponding to number of features selected.

Results for method 1:

Case A: Classification using first six features

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	40.3348
Random Forest Classifier	Macular Edema	59.5789
Support Vector Classifier	Diabetic Retinopathy	43.7639
Support Vector Classifier	Macular Edema	59.2192
Convolutional NN	Diabetic Retinopathy	39.0054
Convolutional NN	Macular Edema	55.8232

Case B: Classification using last six features

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	48.4999
Random Forest Classifier	Macular Edema	63.0653
Support Vector Classifier	Diabetic Retinopathy	45.3214
Support Vector Classifier	Macular Edema	60.2222
Convolutional NN	Diabetic Retinopathy	44.0014
Convolutional NN	Macular Edema	60.2010

Case C: Classification using combination last six features and homogeneity and entropy

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	52.3251
Random Forest Classifier	Macular Edema	66.5841
Support Vector Classifier	Diabetic Retinopathy	52.5411
Support Vector Classifier	Macular Edema	66.8749
Convolutional NN	Diabetic Retinopathy	50.5058
Convolutional NN	Macular Edema	62.6874

Case D: Classification using all features

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	54.474
Random Forest Classifier	Macular Edema	70.147
Support Vector Classifier	Diabetic Retinopathy	50.543
Support Vector Classifier	Macular Edema	66.447
Convolutional NN	Diabetic Retinopathy	53.112
Convolutional NN	Macular Edema	61.447

Case E: Classification after using PCA

Machine Learning Model	Disease	Accuracy	Improvement
Random Forest Classifier	Diabetic Retinopathy	54.578	0.487
Random Forest Classifier	Macular Edema	71.258	1.111
Support Vector Classifier	Diabetic Retinopathy	51.423	0.880
Support Vector Classifier	Macular Edema	66.689	0.242
Convolutional NN	Diabetic Retinopathy	52.858	-0.254
Convolutional NN	Macular Edema	60.639	-0.808

Case F: Classification using all features on large dataset

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	65.574
Random Forest Classifier	Macular Edema	73.407
Support Vector Classifier	Diabetic Retinopathy	61.223
Support Vector Classifier	Macular Edema	68.109
Convolutional NN	Diabetic Retinopathy	68.004
Convolutional NN	Macular Edema	70.147

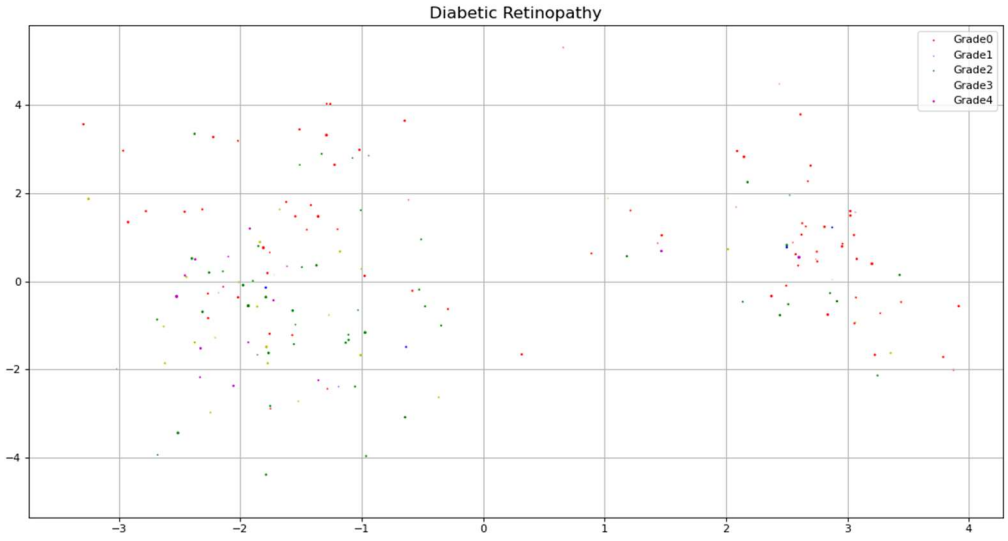
Case G: Classification after using PCA on large dataset

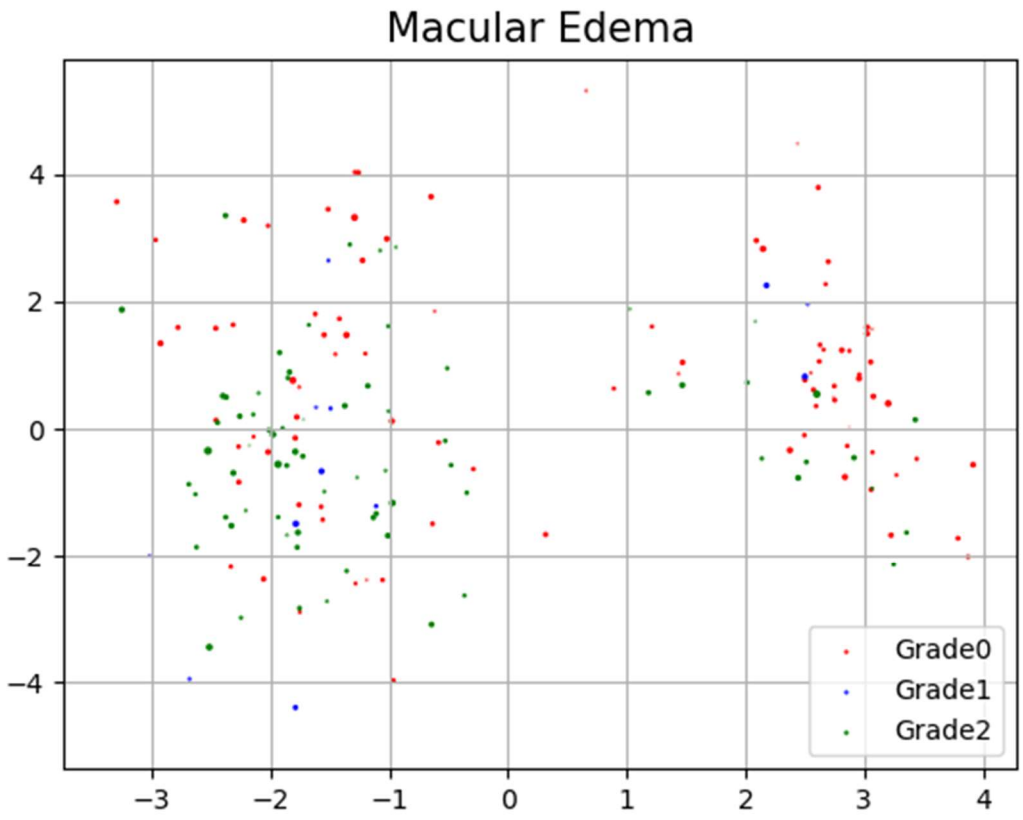
Machine Learning Model	Disease	Accuracy	Improvement
Random Forest Classifier	Diabetic Retinopathy	66.121	0.547
Random Forest Classifier	Macular Edema	73.448	0.041
Support Vector Classifier	Diabetic Retinopathy	61.987	0.764
Support Vector Classifier	Macular Edema	68.274	0.165
Convolutional NN	Diabetic Retinopathy	68.241	0.237
Convolutional NN	Macular Edema	70.522	0.375

Method 2: By using each pixel of image as a feature

Every pixel of an image tells a story about that image, so we used this in implementing our second method.

For an image of m*n pixels, we converted into an array of length m*n. This array was used as a feature vector in image classification model.
We first used this feature vector in our model and then used dimensionality reduction on it to improve our model.





1	0	0	0	0	0	0	0	0	0	0	1	0	2	0	3	1	3	3	3	3	2
0	0	0	0	0	2	2	1	1	0	0	0	2	3	3	2	3	3	3	3	3	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
0	0	0	0	0	1	2	1	3	3	3	3	3	3	3	3	3	3	3	3	3	
0	0	0	3	3	1	1	0	0	1	2	3	3	3	3	3	3	3	3	3	3	
0	0	0	0	0	0	0	1	0	0	0	0	1	1	2	2	0	0	0	1	2	
1	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	
1	0	0	0	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	1	2	
0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	1	1	1	
0	1	2	3	3	2	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	
0	0	0	1	0	0	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	1	2	3	0	0	1	
1	3	3	3	2	2	2	3	3	3	3	3	3	3	3	3	3	2	3	3	3	
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	2	0	1	2	
0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0	
0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	
0	0	0	0	0	1	1	1	0	2	3	2	0	0	0	0	0	1	0	0	1	
0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	2	1	1	2	2	2	
3	0	1	3	1	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
0	0	0	0	0	0	0	0	2	1	0	0	1	1	1	0	2	2	4	0	2	
2	0	0	0	0	1	2	0	1	1	1	2	3	2	1	0	0	0	0	1	1	
0	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	4	3	3	3	3	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Snapshot of dataset using pixel values as feature vector

Results for method 2:

Case A: Classification using all features

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	35
Random Forest Classifier	Macular Edema	38
Support Vector Classifier	Diabetic Retinopathy	37
Support Vector Classifier	Macular Edema	40
Convolutional NN	Diabetic Retinopathy	39
Convolutional NN	Macular Edema	40

Case B: Classification after using PCA with dimension = 200

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	41
Random Forest Classifier	Macular Edema	45
Support Vector Classifier	Diabetic Retinopathy	41
Support Vector Classifier	Macular Edema	43
Convolutional NN	Diabetic Retinopathy	41
Convolutional NN	Macular Edema	43

Case C: Classification after using PCA with dimension = 100

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	41
Random Forest Classifier	Macular Edema	46
Support Vector Classifier	Diabetic Retinopathy	43
Support Vector Classifier	Macular Edema	44
Convolutional NN	Diabetic Retinopathy	41
Convolutional NN	Macular Edema	43

Case D: Classification after using PCA with dimension = 50

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	44
Random Forest Classifier	Macular Edema	49
Support Vector Classifier	Diabetic Retinopathy	47
Support Vector Classifier	Macular Edema	45
Convolutional NN	Diabetic Retinopathy	43
Convolutional NN	Macular Edema	45

Case E: Classification after using PCA with dimension = 25

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	47
Random Forest Classifier	Macular Edema	51
Support Vector Classifier	Diabetic Retinopathy	50
Support Vector Classifier	Macular Edema	48
Convolutional NN	Diabetic Retinopathy	45
Convolutional NN	Macular Edema	46

Case F: Classification using all features on large dataset

Machine Learning Model	Disease	Accuracy
Random Forest Classifier	Diabetic Retinopathy	59.124
Random Forest Classifier	Macular Edema	63.248
Support Vector Classifier	Diabetic Retinopathy	61.332
Support Vector Classifier	Macular Edema	55.104
Convolutional NN	Diabetic Retinopathy	59.584
Convolutional NN	Macular Edema	57.111

Case G: Classification after using PCA on large dataset

Machine Learning Model	Disease	Accuracy	Improvement
Random Forest Classifier	Diabetic Retinopathy	60.111	0.987
Random Forest Classifier	Macular Edema	65.124	1.876
Support Vector Classifier	Diabetic Retinopathy	62.147	0.814
Support Vector Classifier	Macular Edema	58.274	3.170
Convolutional NN	Diabetic Retinopathy	62.241	2.657
Convolutional NN	Macular Edema	60.522	3.411

4. Technologies and Learning:

- Python: Python is an interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, very high-level dynamic data types, and classes. It supports multiple programming paradigms beyond object-oriented programming, such as procedural and functional programming.
- Python Libraries: We used various python libraries like pandas, matplotlib, scikit-learn and NumPy. NumPy is a library consisting of multidimensional array objects and a collection of routines to process them. Pandas is used for high performance data analysis and manipulation. Matplotlib is a popular data visualization package for python. Scikit-learn is useful and robust python library for machine learning and statistical modelling. Keras is open-source neural network library that runs on top of TensorFlow. It is modular, fast and easy to use. Keras is a high-level wrapper API for the low-level API, that runs on Theano, TensorFlow or CNTK. Keras High-Level API handles the way we make models, defining layers, or set up multiple input-output models. In this level, Keras also compiles our model with loss and optimizer functions, training process with fit function. OpenCV for python was used for various image processing technique

- Machine Learning Algorithms: We learnt about various machine learning algorithms that can be used for classification like support vector classifier, random forest classifier, neural networks, k-means algorithm, etc.

5. Final Result

Disease	Best Model with Accuracy	Kaggle Competition Accuracy
Diabetic Retinopathy	CNN => 68.241%	84.957
Macular Edema	Random Forest Classifier => 73.448	84.478

We conclude that our CNN (Convolutional Neural Network) model was giving best accuracy for classifying diabetic retinopathy and our Random Forest Classifier model was giving best accuracy for classifying macular edema.

6. Conclusion:

During this project of 7th semester, we learnt about diabetic retinopathy and the spread of this disease and how many people lose their eyesight due to detection at later stages.

We also learnt about various image processing techniques like histogram equalization, morphological processing, thresholding and edge detection, how to normalize the image to improve the speed of a machine learning algorithm, and how to improve accuracy of our model by various dimensionality reduction techniques.

We were successful in detecting various important features of a fundus image and then developed a classifier system for diabetic retinopathy and macular edema.

7. References:

- *Joshi, Shilpa, and P. T. Karule. "Retinal blood vessel segmentation." International Journal of Engineering and Innovative Technology (IJEIT) 1.3 (2012): 175-178.*
- *Tripathi, Shraddha, et al. "Automatic detection of exudates in retinal fundus images using differential morphological profile." International Journal of Engineering and Technology 5.3 (2013): 2024-2029.*
- *Decencière, Etienne, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, & Jean-Claude Klein.*
- *"FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE." Image Analysis & Stereology [Online], 33.3 (2014): 231-234. Web. 1 May. 2017.*
- *Nixon Aguando, Feature Extraction and Image Processing Book*
- *American Academy of Ophthalmology, Diabetic Retinopathy Guide*
- *Grading diabetic retinopathy (DR) using the Scottish grading protocol, William Wykes, 2015; 28(92): 72-73.*
- *Feature Extraction and Feature Selection from TowardsDataScience.*
- *Machine Learning Cookbook, Chris Alban. O Reilly Publications.*