Deep Learning for Detection and Severity Classification of Diabetic Retinopathy

Anuj Jain
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
anujjaink9@gmail.com

Arnav Jalui
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
arnavjalui@gmail.com

Jahanvi Jasani
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
jahanvijasani.46@gmail.com

Yash Lahoti
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
yashlahoti97@gmail.com

Prof. Ruhina Karani
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
ruhina.karani@djsce.ac.in

***Abstract*—The objective of this project is to automate the detection of Diabetic Retinopathy and evaluate the severity with high efficiency, through an overall feasible approach. This project explores the use of various Convolutional Neural Network Architectures on images from the dataset after being subjected to appropriate image processing techniques like local average color subtraction to aid in highlighting the germane features from a fundoscopy, thereby augmenting the detection and evaluation process of Diabetic Retinopathy as well as serve as an expert guidance system for practitioners around the world.**

***Keywords*—Deep Learning, Diabetic Retinopathy, ConvNets, Image Processing Introduction**

I. Introduction

A. Prior Work and History

There exist multiple techniques for Diabetic Retinopathy (DR) diagnosis, an ocular manifestation of diabetes that affects more than 75% of patients with longstanding diabetes and is the leading cause of blindness for the age group 20-64 [1]. Major Challenging in the ophthalmology research is process of automation which is useful in prior detection and diagnosis of advance eye diseases[2]. In this paper, we focus on diagnosis through the use of retinal fundus images, which involves careful examination of photographs of the retina taken with expensive equipment by trained clinicians. This detection technique is very resource intensive and requires very specialized clinician knowledge [3]. There have been many advancements in the development of algorithms and morphological image processing techniques that extract features prevalent in patients with diabetic retinopathy. For an overview of such algorithms consult [4]. Faust et al. [5] provide a very comprehensive analysis of models that use explicit feature extraction to DR screening. Limitations of these studies are in the magnitude of their scope, the homogeneity of the dataset and, the narrowness of the features that have been extracted from the images. Vujosevic et al. [6] build a binary classifier on a dataset of 55 patients by explicitly forming single lesion features. The authors in [7] use morphological image processing techniques to extract various features of DR and then train a Support Vector Machine (SVM) on a data set of 331 images achieving sensitivity 82% and specificity 86%. The authors in [8] report the accuracy of 90% and sensitivity of 90% (on binary classification task with a dataset of 140 images) using image processing techniques to extract the area of blood vessels, the area of exudates, and texture features which are then fed into a small Neural Network. Recent work by Rahim et al. uses fuzzy image processing techniques (fuzzy histogram equalization and fuzzy edge detection) for a DR detection system. A thorough analysis of various other DR detection methods can be accessed [9].

B. Background Literature Survery

Considerable developments have been made in algorithm generation and morphological image processing techniques for extracting features of Diabetic Retinopathy from the fundus inputs provided. There exists heterogeneity in the sense that there is conflict during the selection of training method of the convolutional neural networks.

A dilemma between the use of transfer learning for training the models or to train the models from scratch was observed. Once trained, test images are passed forward through the network and the model attempts to predict the severity of diabetic retinopathy. The grading between the ordinal classifications may be subtle and variable between even expert graders. Some works consisted of a panel of expert ophthalmologists to create standards for retinopathy grading as described previously.

The work by Walter et. al. [4] describes previous feature centric algorithms and outlines successful algorithms for optic disk segmentation and exudate detection. The review by Faust et. al. covers additional algorithms for segmenting haemorrhages and evaluating texture, as well as reviewing the sensitivity and specificity of many classification methods (consisting of feature centric algorithms). Preliminary readings of the literature suggest that most previous works focused on using hand designed features to train machine learning algorithms.

While these algorithms are suitable for clean high resolution images, in practice, they encounter difficulties with artefacts and low image quality, which is a very potent problem when dealing with real life imagery of the retina. We need to factor for the reality that not all images that will be used to train the system and will be passed through the system for prognosis will be ideal and most of the images will contain some undesirable commodities that will hinder the performance of an otherwise perfect system.

One Shortcoming of these studies are in the diversity of their scope (all the studies present results derived from one or few phenotypes). This homogeneity of the dataset and, the narrowness of the explicit features extracted from the images render it inefficient in Asian countries like India. In addition to understanding and reviewing the technical aspects of previous works done on DR detection, we conducted an extensive research on the medical side of the project from various medical journals and online databases.

The attribute of the proposed system that sets it apart from the similar other projects of past is its focus on implementing a DR detection system in rural areas which will act as an automated and cheaper alternative to the conventional method of manual diagnosis by experts that involves considerable traveling and is prone to waste valuable time in providing results to the patients.

To create a system that is available, affordable and feasible for the remotest parts of our nation we studied the prevalence of DR in the rural population as well as the common problems faced by them during the diagnostic phase [7]. Multiple surveys conducted by Sankara Nethralaya suggested that,

Nearly 1 of 10 individuals in rural South India, above the age of 40 years, showed evidence of type 2 diabetes mellitus. Likewise, among participants with diabetes, the prevalence of diabetic retinopathy was around 10%; the strongest predictor being the duration of diabetes [10].

Another study focused on understanding the disparity in the necessary medical facilities present in different parts of country showed that,

As of 2012, there are 655 ophthalmologists who are registered in the Vitreoretinal Society of India. In 2015, based on the Novartis India user data, 1058 ophthalmologists inject intravitreal ranibizumab across India. This number is far less to tackle the projected load of DR in India. Against a national ophthalmologist: Population ratio of 1:107,000, there are certain regions in India which have a ratio of 1:9000 while in other regions there is only one ophthalmologist for 608,000 population [11].

Based on the further study of other such surveys and research revealed to us a glaring lack of standard medical technology in the less developed parts of India [7]. Reviewing these types of medical research studies and social surveys provided us with a concrete and fundamental understanding of the shortcomings in medical infrastructure of rural areas that we need to counter through our proposed system.

Marco Alban and Tanner Gilligan of Stanford University presented a paper – "Automated Detection of Diabetic Retinopathy using Fluroescein Angiography Photographs", which served as a technical and conceptual basis for the implementation aspects of our project.

They presented an analysis of a model for multi-class identification of the severity of DR from fluorescein angiography photographs. The model had performed well in comparison to human evaluation metrics. It could be further developed through exploration more nuanced data normalization and denoising techniques. For instance, apriori knowledge of the sources of error for equipment used to capture fundus photographs could help in building more robust normalization schemes. Other work may be in combining weak learners in ensembles or using an ensemble of classifiers trained on raw image pixels and trained on explicit feature extractors Dataset

C. EyePACS

We used the images from the EyePACS dataset provided as a part of the competition hosted on Kaggle. There existed a total of 35126 images in the EyePACS dataset, which were of varied sizes and were classified into predetermined stages of severity of the disease. These predetermined stages are as follows,

- No DR (Class 0)
- Mild DR (Class 1)
- Moderate DR (Class 2)
- Severe DR (Class 3)
- Proliferative DR (Class 4)

The "Mild", "Moderate" and "Severe" severity levels are non-proliferative versions of Diabetic Retinopathy, wherein the blood vessels in the retina are blocked in increasing order, respectively [12].

The dataset contains images of the left as well as the right retina. The 35126 images present in the dataset are captured with different models and types of cameras, which can affect the visual appearance of left vs. right. Some of the images are depicted as one would see the retina anatomically (macula on the left, optic nerve on the right for the right eye). Others are shown as one would see through a microscope condensing lens (i.e. inverted, as in a typical live eye exam). There are generally two ways to tell if an image is inverted:

i. It is inverted if the macula (the small dark central area) is slightly higher than the midline through the optic nerve. If the macula is lower than the midline of the optic nerve, it's not inverted.

ii. ii. If there is a notch on the side of the image (square, triangle, or circle) then it's not inverted. If there is no notch, it's inverted.

It consists of 25810 images of Class 0, 2443 of Class 1, 5292 of Class 2, 873 of Class 3, 708 of Class 4.

II. Implementation

A. Dataset Preparation

The images in the dataset were split into training, validation and test sets with 60% of the total images in the dataset being used for training the models, 20% used for validation purposes and the remaining 20% being used for testing the trained models. As mentioned earlier, the images were of different dimensions. For the purpose of training a neural network, it is necessary that all the images be of a standard size.

1) Normalization: The images were centered and cropped to a size of 256×256 pixels as this resolution was found to retain almost all of the features offered by a fundus image while being small enough to reduce training overheads on the neural network. There existed a wide range of variations present in the images with respect to the brightness, contrast, color, presence of lens glare, the presence of noise, etc. It became imperative to perform image normalization procedures on the images to achieve some level of uniformity in the images, which would consequently improve training speeds and accuracy. Theoretically, severe DR leads to higher heterogeneity than mild, moderate or no DR in a

fundus photograph [13]. We subtracted the local average color from the images to normalize them.

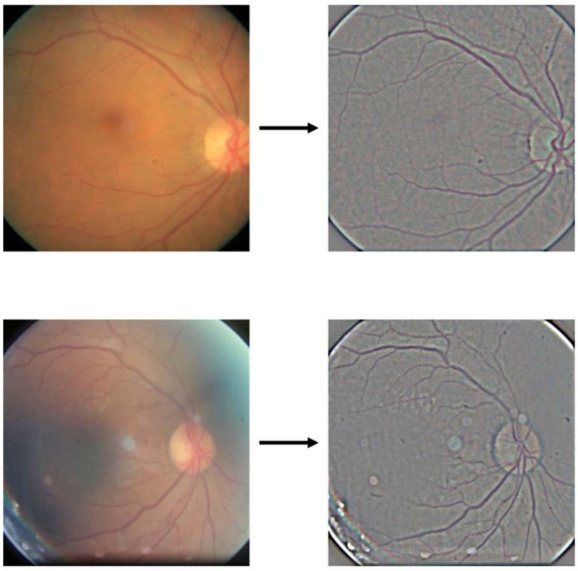

Fig. 1. Two images from the Dataset. Original image on the right and Normalized images on the left

2) Data Augmentation: We observed during training that after a number of epochs, the model was starting to over-fit the dataset. Training accuracy was increasing by the epoch but the same could not be said about validation accuracy. Also, there existed an imbalance in the classes of data in the dataset. The class 0 was about 75% of the entire dataset and the rest of the classes were about 25% of the dataset. This situation warranted for use of techniques that mitigate the issue of over-fitting as well as data imbalamce. We rotated the images in the training set by 90 and 270 degrees to increase the count of training images significantly.

TABLE I
NUMBER OF IMAGES PER CLASS OF TRAINING DATASET

Classes	Original Data	Augmented Data
0	15534	15534
1	1461	4383
2	3140	9420
3	522	1566
4	418	1254

Thus, the ratio of Images without DR to Images with DR increased from 2.8:1 to 1:1.07 after augmentation of images in the training dataset.

B. Convolutional Neural Networks

Convolutional Neural Networks (CNN) have time and again proven to be the most effective neural network architectures for image classification and object detection. The task of identifying the presence of diabetic retinopathy in a fundoscopy and evaluating the severity of the disease is also best performed by CNN. The selection of a CNN architecture and fine-tuning it's parameters to obtain optimal results was the primary task at hand. We chose to experiment with a few established and widely renowned pre-trained networks, viz. the VGG16, VGG19 and Inception v3. The top-most layers were discarded and replaced with something that suited our needs. We added two fully-connected layers at the top of the architecture, the former layer consisting of 4096 neurons and the latter consisting of 1024 neurons. They used the Rectified Linear Unit (ReLU) activation function. The top-most layer in each case is a five neuron layer, where one neuron will be fired for each of the predetermined stages of the disease. This layer makes use of the Softmax activation function. All the architectures were implemented with the help of Keras framework, running on top of a Tensorflow backend. Training was conducted by freezing the weights of the convolutional layers as well as by including these layers in the training process as well.

1) VGG16: The first model that we used was the VGG16 architecture proposed by Karen Simonyan and Andrew Zisserman in [14]. It used 3×3 convolutional kernels and 2×2 pooling kernels. Maxpooling was used in the pooling layers. There were two fully connected layers at the top of the model, each consisting of 4096 layer. The topmost layer was the predictions layer, which consists of 1000 nodes to indicate the prediction of 1000 objects [14]. We altered the final two layers to suit our needs. The ultimate layer in our implementation of VGG16 has 1024 nodes in the fully connected layer and the predictions layer has 5 nodes instead of 1000 nodes to predict the 5 different classes of Diabetic Retinopathy.

2) VGG19: The next model that we used for our project was the VGG19 architecture also proposed in [14]. The main difference between VGG16 and VGG19 in terms of network architecture is the additional convolutional layer in the 3rd, 4th and 5th convolutional blocks. We altered the VGG19 architecture on lines similar to the VGG16 architecture. The penultimate fully connected layer had 1024 nodes and the topmost predictions layer had 5 nodes for predicting 5 levels of Diabetic Retinopathy.

3) InceptionV3: The Inception architecture is one of the newer CNN architectures proposed in [15]. With a depth of 42 layers, the computation was much more efficient than any of the VGGNets. All the convolutional layers are 3×3. The max-pooling layers were 3×3 as well 8×8.

III. RESULTS

Our models were run on Google Colab with a Tesla K80 GPU (12 GB) and on a local machine with Nvidia GTX 1050 Ti GPU (4 GB). All models were developed by using Keras deep learning framework on a Tensorflow backend.

A. Metrics

We used 4 evaluation metrics for evaluating the performance of the trained models, viz. Accuracy, Sensitivity, Specificity and Precision. Metrics like accuracy were calculated on a per class basis (a total of 5 classes) and the rest were calculated on a 2-class basis, i.e., based on presence or absence of diabetic retinopathy in a retinal fundus images.

TABLE II
CONFUSION MATRIX FOR VGG16

	2 Class Predictions	
	Predicted Non DR	***Predicted DR***
Actual Non DR	4825	292
Actual DR	1516	393

TABLE III
2 CLASS METRICS FOR VGG16

Sensitivity	0.206
Precision	0.574
Specificity	0.943
Accuracy	0.743

TABLE IV
CONFUSION MATRIX FOR VGG19

	2 Class Predictions	
	Predicted Non DR	***Predicted DR***
Actual Non DR	4952	165
Actual DR	1215	694

TABLE V
2 CLASS METRICS FOR VGG19

Sensitivity	0.364
Precision	0.808
Specificity	0.968
Accuracy	0.804

TABLE VI
CONFUSION MATRIX FOR INCEPTION V3

	2 Class Predictions	
	Predicted Non DR	***Predicted DR***
Actual Non DR	4286	831
Actual DR	912	997

TABLE VII
2 CLASS METRICS FOR INCEPTION V3

Sensitivity	0.522
Precision	0.545
Specificity	0.838
Accuracy	0.752

Table II to Table VII show the results for 2 classes i.e. with and without Diabetic Retinopathy. Comparatively, Inception V3 provides better overall results.

TABLE VIII
CONFUSION MATRIX FOR VGG16

	3 Class Predictions		
	Predicted Non DR	***Predicted NPDR***	***Predicted PDR***
Actual Non DR	4825	276	16
Actual NPDR	1446	277	35
Actual PDR	70	38	43

TABLE IX
3 CLASS METRICS FOR VGG16

	Sensitivity	***Precision***
Non DR	0.943	0.761
NPDR	0.158	0.469
PDR	0.245	0.457
Average	0.462	0.562

TABLE X
CONFUSION MATRIX FOR VGG19

	3 Class Predictions		
	Predicted Non DR	***Predicted NPDR***	***Predicted PDR***
Actual Non DR	4952	156	9
Actual NPDR	1177	563	18
Actual PDR	38	65	48

TABLE XI
3 CLASS METRICS FOR VGG19

	Sensitivity	***Precision***
Non DR	0.968	0.803
NPDR	0.320	0.718
PDR	0.318	0.640
Average	0.535	0.720

TABLE XII
CONFUSION MATRIX FOR INCEPTION V3

	3 Class Predictions		
	Predicted Non DR	***Predicted NPDR***	***Predicted PDR***
Actual Non DR	4286	805	26
Actual NPDR	888	837	33
Actual PDR	24	61	66

TABLE XIII
3 CLASS METRICS FOR INCEPTION V3

	Sensitivity	***Precision***
Non DR	0.838	0.825
NPDR	0.476	0.491
PDR	0.437	0.528
Average	0.584	0.615

The remaining tables (table VIII to table XIII) depict the results obatined for 3 classes i.e. No DR, Non-Proliferative DR(NPDR) and Proliferative DR(PDR).

Table XIV to XVI depict the sensitivity and precision for the models in a 5-class division. These are the final results obtained through training. VGG16 provided us with an accuracy of 71.7%, whereas the same for VGG19 76.9% and Inception v3 was 70.2%. Even though the accuracy for Inception v3 is less, it is a better detector for case which actually contain DR and does not pass of positives as false negatives.

B. Hyperparameters

We selected the crucial hyperparameters like learning rate by leveraging standard practices of increasing and decreasing the value of learning rate by a factor of log10. We brought into use some of the widely used optimization techniques such as Stochastic Gradient Descent (SGD), Root Mean Square Propagation (RMSprop) and Adaptive Moment Estimation (Adam). We found out that SGD was providing us with consistent and desirable results, thus all of our models used for the final evaluation were trained using SGD.

TABLE XIV
5 Class Metrics for VGG16

	Sensitivity	Precision
0	0.943	0.761
1	0.012	0.048
2	0.133	0.366
3	0.102	0.310
4	0.285	0.458
Average	0.295	0.387

TABLE XV
5 Class Metrics for VGG19

	Sensitivity	Precision
0	0.968	0.803
1	0.002	0.333
2	0.309	0.538
3	0.345	0.399
4	0.318	0.640
Average	0.388	0.543

TABLE XVI
5 Class Metrics for Inception v3

	Sensitivity	Precision
0	0.838	0.825
1	0.031	0.070
2	0.447	0.381
3	0.401	0.353
4	0.437	0.538
Average	0.431	0.431

C. Model Performance

By using 3 different CNN architectures, we found out that for the problem statement at hand, the performance of the model was directly linked to the number of convolutional and pooling layers in the CNN. Elaborating further, this means that the results obtained on the test set by running it through the VGG19 model eclipsed the results obtained by VGG16. Furthermore, the Inception architecture outdid both the VGG16 and VGG19 architectures in terms of overall performance on the test set. The training time was also directly linked to the number of layers in the CNN, but a gain in model performance outweighed the additional time cost of model training in the Inception architecture. The results of evaluating the test set of 7026 images belonging to 5 different classes on three different CNN architectures are elaborated in Fig. 2 and Fig. 3

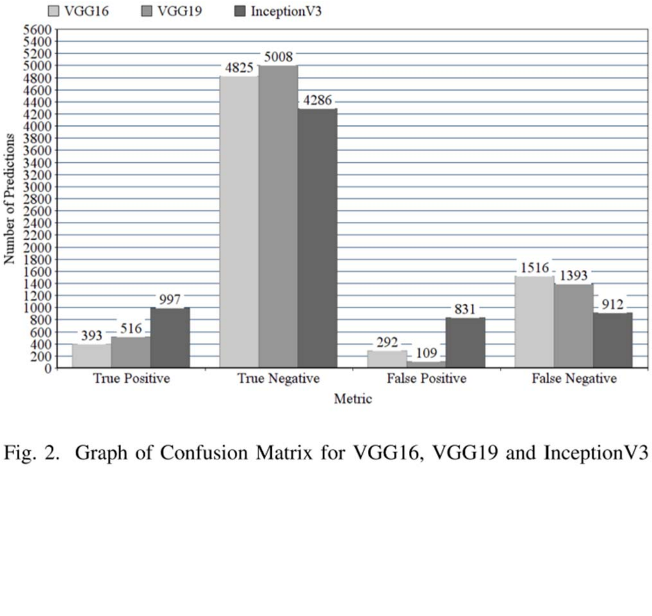

Fig. 2. Graph of Confusion Matrix for VGG16, VGG19 and InceptionV3

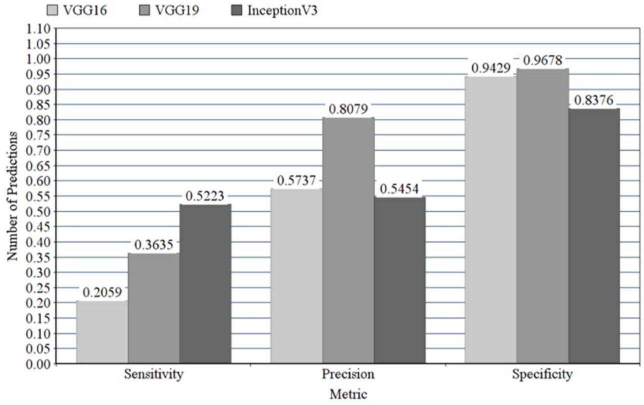

Fig. 3. Graph of Sensitivity, Precision and Specificity for VGG16, VGG19 and InceptionV3

IV. References

[1] Linda Geiss et al. Engelgau, Michael. The evolving diabetes burden in the united states. Annals of Internal Medicine, 2004.

[2] Zhang B, Karray F, Li Q, Zhang L. Sparse Representation Classifier for micro aneurysms detection and retinal blood vessel extraction. Inform Sci. 2012; 200: 78-90. doi: 10.1016/j.ins.2012.03.003

[3] Grading diabetic retinopathy from stereoscopic color fundus photographs an extension of the modified airlie house classification: Report number 10

[4] P. Massin A. Erginay T. Walter, J. Klein. A contribution of image processing to the diagnosis of diabetic retinopathy detection of exudates in color fundus images of the human retina. IEEE Transactions on Medical Imaging.

[5] K. Ng J. Suri O. Faust, R. Acharya. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review. Springer Science, Journal of Medical Systems.

[6] Benetti E. Massignan F. Pilotto E. Varano M. Cavarzeran F. Avogaro A.Vujosevic, S. and E. Midena. Screening for diabetic retinopathy: 1 and 3 non mydriatic 45-degree digital.

[7] Ng EY Chee C Tamura T. Acharya U, Lim CM. Computer-based detection of diabetic retinopathy stages using digital fundus images. Proceedings of the Institute of Mechanical Engineers, 545-553, 2009.

[8] Bhat P. S. Acharya U. R. Lim C. M. Nayak, J. and M Kagathi. Automated identification of different stages of diabetic retinopathy using digital fundus images.

[9] Hykin PG Fraser-Bell S, Kaines A. Update on treatments for diabetic macular edema. Current Opinion in Opthalmology, 2008.

[10] Raman R., Ganesan S., Pal S., Kulothungan V., Sharma T. Prevalance and risk factors for diabetic retinopathy in rural India. Sankara Nethralaya Diabetic Retinopathy Epidomology and Molecular Genetic Study III

[11] Rajiv R., Laxmi G., Sangeetha S., Tarun S. Diabetic retinopathy: An epidemic at home and around the world. Symposium- Diabetic Rtinopathy in India

[12] Vinod Patel Eva M Kohner and Salwan M B Rassam. Role of blood flow and impaired autoregulation in the pathogenesis of diabetic retinopathy. American Diabetes Association.

[13] Gen-Min Lin, Mei-Juan Chen, Chia-Hung Yeh, et al., “Transforming Retinal Photographs to Entropy Images in Deep Learning to Improve Automated Detection for Diabetic Retinopathy,” Journal of Ophthalmology, vol. 2018, Article ID 2159702, 6 pages, 2018. https://doi.org/10.1155/2018/2159702.

[14] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv.org, 2015

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. arXiv.org, 2015