AUTOMATED DETECTION AND GRADING OF DIABETIC RETINOPATHY USING CNN

Yash Choudhary¹, M. Eliazer², Aditya Jain³
¹ys7613@srmist.edu.in, ²eliazerm@srmist.edu.in, ³av9427@srmist.edu.in
¹,²³Computer Science and Engineering Department
SRM Institute of Science and Technology
Chennai, India

ABSTRACT

Diabetic Retinopathy (DR) is quite a major threat when complications of diabetes are taken into account which might result in permanent blindness if not treated within a stipulated time. It is more common in working-age people. Research says that close to 4 million people with diabetes get blind each year worldwide by DR. Early diagnosis are considered an effective way to detect and endure such a difficulty. DR detection method can be classified on the categories of feature extraction method as traditional feature extraction which relies on machine learning(ML) and Deep Convolutional Neural Network(CNN). The primary objective of our research is to develop an effective ML/DL technique to investigate the DR images captured by retinal camera for autonomous DR detection. The study says that the traditional feature extraction methods are less accurate than some other approaches practiced. We try to create a CNN based model for fast and accurate DR detection.

Keywords: Feature Extraction, Deep Convolutional Neural Network (CNN), Severity scale, Fundus.

I. INTRODUCTION

Diabetic Retinopathy (DR) is a severe complexity of diabetes leading to permanent blindness if left untreated. DR is classed into four standard types, mild Non-Proliferative Diabetic Retinopathy (NPDR), moderate and severe NPDR, and lastly Proliferative DR (PDR). It's a problem in the retina that majorly happens when blood vessels in the retina are hampered due to various reasons.

It is said that if someone has diabetes, it is quite important to get a detailed examination of the dilated eye at least once a year. In the starting people have trouble reading or seeing faraway objects. Signs of DR can be bleeding of the blood vessels of retina into vitreous. Some small red spot-scan also appear on blood vessels of the retina. Hemorrhage can be formed if early walls of capillaries are broken when not treated well. Permanent vision loss can be caused if exudates appear on the retina. Diabetic Retinopathy is growing at a faster rate and has affected near about 30% of the population.

Diagnosis of DR is performed in the clinic by experts with the help of accurate fundus image capturing devices. Post capturing the retinal images, techniques, like fundus photography and tomography are used.

Expensive designand usage costs are major challenges included in these techniques. Experts use these techniques and after that an ophthalmologist reviews and diagnoses the fundus image.

It takes some days for an ophthalmologist to do the retinal image diagnosis. Affected patients living in rural areas suffer from delayed diagnosis and necessary treatment because of costly diagnosing equipment and few ophthalmologists and centers. It’s a thorough and time-consuming job for an ophthalmologist to examine the entire image and analyze lesions because the size of a lesion instance is quite small, which makes it difficult to detect.

To strengthen the DR detection in early screenings it is important to develop an efficient tool which brings out maximum accuracy. Considering the growth an automatic grading-based screening system will be helpful.
Automated grading system is cost-effective compared to manual grading. To chop back the time and price of the screening, automated algorithms are developed to research the images acquired.

The analysis which used the deep CNN method is extremely particular during an automated DR severity grading and also very useful during a large-scale database. Feature enhancing methods like matched filtering and region growing are combined with SVM and Neural networks and is utilized for various classification problems of DR severity. There has been growth of Graphics Processing Units (GPUs) because of growth of a Deep CNN based feature extraction method. As a result, deep learning models for DR detection is popular among several researchers. Deep learning ensures high accuracy and performance and is a powerful tool for DR detection.

This paper also provides insights on the severity grading stage of these methods. Quality type for the severity capacity that supports the lesions existing in an exceeding retina is described in Fig1.

A Computer Vision Model can take these images and classify them as DR or no DR. The model will grade DR images on a scale of 0-4 severity according to the standards. This suggests a report and helps refer the patient to the ophthalmologist in case of abnormalities.

The model practices a detailed three-step pipeline that involves capturing retinal fundus images, image preprocessing and augmentation, and classification and grading of the images. During this paper a smartphone can be used for data collection and a trained model for detection helps to diagnose DR instantly at a patient's place.

The below figure gives visual reference of how a normal retinal differs from a severe NDPR retina and other stages in-between.

![Fig. 1. Human retina and NPDR stages: (a) normal retina (b) mild NPDR, (c) moderate NPDR and (d) severe NPDR.](image)

**II. RELATED WORKS**

The capability of CAD systems has grown immensely and it has been seen that to accurately it detects the grades of DR accurately and is now famous among several researchers. In the last few years, various research work focused on the development of Computer-Aided systems to automatically detect DR were recorded. The papers surveyed by us used these datasets: EYEPACS, MESSIDOR, DRIVE, STARE, DIARETDB, KAGGLE and SIDPR.

Yehui Yang et al. in their paper presented a robust framework which could collaboratively utilize both patch-level lesion and image level annotations. The proposed algorithms in this paper showed favorable performance. Their lesion attention Generator could be pre-trained with patches and could avoid the missing label problem. The model presented could also detect lesion in an image in one go. It was also seen that there was still a need to detect more types of lesion precisely[8].

www.turkjphysiotherrehabil.org
Gwenole Quellec et al. in their paper trained end to end with image supervision only which is just like black box AI algorithms. The foreground or background separation was trained through self supervision. The model also offered high performance and explain-ability which helped in facilitating AI deployment. But it was seen that the model was less focused on creating an efficient and accurate solution[17].

Yi Zhou et al. in their paper proposed and used a large fine-grained annotated DR dataset, called FGADR[7]. It was also proved that joint classification and segmentation methods have a better performance rate. An inductive transfer learning method was also developed. It was also mentioned that there were some difficulties differentiating between grade 3 and 4 precisely. The approach used in this paper was quite close to the perfect solution so there were chances of exploration in the same direction.

Saket S. Chaturvedi et al. in their paper presented a DenseNet121 model[4]. They also employed several modifications and preprocessing to improve performance of the model. The model presented was good enough to be employed in clinics for DR Detection. The major point of concern was that it took lots of efforts in terms of data – processing.

Misgina Tsighe Hagos et al. in their paper used a CNN method with a binary tree-based ensemble of classifiers to increase the performance. The model trained could provide a usable point-of-care diagnostic services for DR[1]. The model was so well equipped that it could even be used with other medical applications. The developed application could only diagnose images from a phone’s gallery.

Hangwei Zhuang et al. in their paper used transfer learning to train deeper models which increased the accuracy[2]. They proposed two different solutions regarding Diabetic Retinopathy (DR) classification. Within the first approach, a shallow neural specification is introduced. This model performs well on the classification of the foremost frequent classes while fails at classifying the less frequent ones. Within the second approach, transfer learning is employed to retrain the last modified layer of the deep neural network to enhance the generalization ability of the model to the less frequent classes. It was also seen that the method used could not generalize well due to use of very deep models also categorized as overfitting.

Deep learning (DL) techniques have achieved encouraging outcomes in computer vision problems[14], especially within the field of medical image analysis[15][16]. P. Prentaˇsici and S. Lonˇcaric worked on designing the CNN with several layers and trained it to detect the exudate of the RGB color retina images[12].

Khojasteh et al. compared the accuracy performance of different methods of Deep Learning for exudate detection, which obtained an accuracy of 89.1%[13]. All literature indicates that deep learning and CNN are promising and have a stronger future for automatic diabetic diagnosis supported colored retina images.

Niloy Sikder et al. developed an entire pipeline for early detection of DR using the process of ensemble learning. Their pipeline includes 4 major steps - Image preparation, Image preprocessing, Feature Extraction and Image classification[9]. They used APTOS dataset for experiments. The results weren’t very promising, but it shows that the standard approaches are still valid and may compete with the new advances.

Borys Tymchenko et al. proposed the multi-stage transfer learning approach and an automated method for detection of the stage of DR using single photography of the fundus[5]. They combined various CNN processes which were EfficientNet-B4, EfficientNet-B5, SE-ResNeXt50.

Farzan Shenavarmasouleh and Hamid R. Arabnia proposed a specific approach that used CNNs and transfer-learning to detect/segment two main varieties of lesion, namely, Exudates and Micro-aneurysms that cause DR[21].

Other researchers also tried to create transfer learning with CNN architecture. Hagos et al. tried to coach Inception Net V3 for 5-class classification and pre-trained it on ImageNet dataset achieving an accuracy 90.9%.

III. PROPOSED SOLUTION

3.1 Overview

After lots of research by experts it is said that DR should be detected in the early stages because that way it reduces the severity of DR by manifolds. To strengthen the DR detection in early screenings it is important to develop an efficient tool which brings out maximum accuracy. Considering the growth an automatic grading-

www.turkjphysiotherrehabil.org
based screening system will be helpful. In this paper we aim to develop a Computer Vision Model to detect DR and grade it based on the severity level.

We aim to grade on the scale of 0-4 which covers stages from No DR to Proliferative DR. We aim to develop a model that diagnoses DR at a faster rate by examination of different category of retinal images. To create this model, we will be following a 3-step pipeline: Capturing retinal images, preprocessing the images and augmentation and finally classifying and grading of the segmented images. A smart phone device having a mobile image capturing camera will be efficient to diagnose DR in any area. This can reduce the time constraint challenges of the conventional methods. Considering this we also aim to create a mobile or web application to bundle all the functionalities in a user-friendly manner.

3.2 Dataset

In our research, we have used the dataset from two Kaggle Competitions. These are Diabetic Retinopathy Detection 2015 [24] and APTOS 2019 [23] dataset. The main motivation behind using these was that they were readily available and the dataset was from a dependable source. Also, the quantity of the data was sufficient to back any research. The datasets were joined and re-split according to requirements. 2015 training and test dataset were combined with 2019 train dataset to create the dataset. So the total dataset has 142,796 images, split into 5 grades from 0 to 4. These images were of variable sizes, which we handled during preprocessing.

<table>
<thead>
<tr>
<th>DR Grade</th>
<th>Grade Name</th>
<th>Total Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No DR</td>
<td>60433</td>
</tr>
<tr>
<td>1</td>
<td>Mild DR</td>
<td>5917</td>
</tr>
<tr>
<td>2</td>
<td>Moderate DR</td>
<td>12737</td>
</tr>
<tr>
<td>3</td>
<td>Severe DR</td>
<td>2052</td>
</tr>
<tr>
<td>4</td>
<td>Proliferative DR</td>
<td>1988</td>
</tr>
</tbody>
</table>

Table 1. Grade-wise Dataset Split-up

Fig. 2. Sample Images from dataset. (A) No DR, (B) Mild DR, (C) Moderate DR, (D) Severe DR, (E) Proliferative DR.

3.3 Data Preprocessing

The preprocessing was done in multiple steps. The initial dataset was standardized and augmented to make it consistent and remove class imbalance as much as possible. Also, augmentation makes training more robust and fool-proof. Standard image properties were:

\[
\text{IMG\_SIZE} = 224
\]

\[
\text{NB\_CHANNELS} = 3
\]
After this, the images were processed to remove the black parts as much as possible and enhanced to bring out the parts of the image that were essential to improve training. These preprocessing steps include:

1. Standardize pixel values (Image Normalization)
2. Crop the black parts around the image
3. Enhance the image (Ben Graham’s Method)

After this preprocessing, data of class 1-4 was augmented to remove class imbalance. The augmentations used were:

1. Horizontal Flip (p=1)
2. Vertical Flip (p=1)
3. RandomRotate90 (p=0.5)

3.4 Network Architecture

For our model we used Google’s EfficientNet[25]as the base model. The model was tweaked using some extra layers to improve its capability to generalize. The base model was headless so we added an output layer with “SOFTMAX” activation function to get the output.

Reason behind using EfficientNet [25] is that it is proved to have worked better than other models which have 2-3 times more parameter and take subsequently more time and space to train and run. Our aim was to create a model as light as possible so that it could be deployed in real world situations with almost real-time speed of inference. Due to this we tried multiple different versions of EfficientNet [25] to find the one with best accuracy/speed to size/time ratio. We trained EfficientNet B1, B3 and B5 [25] with variable hyperparameters to find the one that will best suit our cause.

The table and figure below show the proposed model architecture. Our model is a slightly modified and more efficient version of EfficientNetB3 [25].

<table>
<thead>
<tr>
<th>Layer (Type)</th>
<th>Output Shape</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficientnet-b3</td>
<td>(None, 7, 7, 1536)</td>
<td>10608936</td>
</tr>
<tr>
<td>Global Average Pooling</td>
<td>(None, 1536)</td>
<td>0</td>
</tr>
<tr>
<td>Dropout</td>
<td>(None, 1536)</td>
<td>0</td>
</tr>
<tr>
<td>Dense_1</td>
<td>(None, 5)</td>
<td>7685</td>
</tr>
<tr>
<td>Dense_2</td>
<td>(None, 1)</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2. Proposed Network Architecture

3.5 Training
The training was done with single-label classification for the grading of DR considering the 5 standard severity levels. The data was split into train-test-validation sets using the same seed to keep all experimental hyperparameters constant. We trained EfficientNetB1 with batch size 16, B3 with batch size 8 and 16 and B5 with batch sizes 8 and 16 to find out the configuration that works best.

The metrics used to evaluate the model was Quadratic Weighted Kappa. All models were trained for 15 epochs in different configurations to get a comparable and devisable result.

IV. RESULTS

All trainings and calculations were done on Kaggle kernel having 4 CPU cores and Nvidia P100 GPU with 16GB V-RAM. Model was evaluated on a particular set taken from the dataset and same preprocessing steps are training were applied to images at the time of testing.

The main metric used for evaluation is Quadratic Weighted Kappa. “The Kappa coefficient is a chance-adjusted index of agreement. In machine learning it can be used to quantify the amount of agreement between an algorithm’s predictions and some trusted labels of the same objects. Kappa starts with accuracy - the proportion of all objects that both the algorithm and the trusted labels assigned to the same category or class. However, it then attempts to adjust for the probability of the algorithm and trusted labels assigning items to the same category "by chance." This metric typically varies from 0 (random agreement between raters) to 1 (complete agreement between raters). In the event that there is less agreement between the raters than expected by chance, the metric may go below 0.”[22]

\[
\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}
\]

4.1 Performance Evaluation on Severity Grading

The model was used to find predictions on the test set which contained 14280 retina images. These results were used to calculate Precision, Recall and F1-Score which are used to tell how good the model is performing per class. It accounts for class imbalance and is a better metric than plane accuracy.

\[
Recall = \frac{True \ Positives}{True \ Positives + False \ Negatives}
\]

\[
Precision = \frac{True \ Positives}{True \ Positives + False \ Positives}
\]

\[
F1 \ Score = 2 \times \frac{Precision \times \ Recall}{Precision + Recall}
\]

<table>
<thead>
<tr>
<th>Classes</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>No DR</td>
<td>0.85</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>Mild DR</td>
<td>0.42</td>
<td>0.45</td>
<td>0.43</td>
</tr>
<tr>
<td>Moderate DR</td>
<td>0.90</td>
<td>0.82</td>
<td>0.86</td>
</tr>
<tr>
<td>Severe DR</td>
<td>0.81</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>Proliferative DR</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>Macro Avg</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>Weighted Avg</td>
<td>0.81</td>
<td>0.80</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 3. Precision, Recall and F1 scores for EfficientNet B3[25]
The final Quadratic Weighted Kappa scores for Train, Validation and Test sets came out to be 96.8%, 91.02%, 90.28% respectively. These scores can be optimized using thresholding and Grid Search to be 91.25% for validation and 90.55% for test set.

4.2 Confusion Matrix

A confusion matrix compares true labels with predicted labels to find and explain the performance of the classification model over a validation set. As shown in the below table, our model performs well on all classes and exceptionally well on No DR class. It sometimes confuses between No DR and Mild DR but that is normal because even human experts sometimes are not able to classify them correctly.

![Image of confusion matrix](image)

Table 4. Confusion Matrix: Diagonal elements represents the number of correct predictions in that class.

4.3 Accuracy and Loss

We trained our model for 52 epochs. This was decided based on the accuracy and loss curves plotted from the model training history. The curves started flattening at around 35 epochs and almost flattened out completely by 50 epochs. Below are the training accuracy and loss curves for EfficientNet B3 [25]. The curves have plots for loss, validation loss, accuracy and validation accuracy. As evident from the below curves, our model was training well using the mentioned settings. The losses are decreasing and accuracy is increasing.

![Image of accuracy and loss curves](image)

Fig. 3. Accuracy and Validation Accuracy graph

![Image of loss and validation loss curves](image)

Fig. 4. Loss and Validation Loss graph
In this paper we tried to find a model that has both speed and accuracy so that it could be used as an aid to ophthalmologists. It should be fast enough that it does not hinder work and it should be accurate enough to be dependable. In our research we found out that despite common belief that a larger model with more parameters is bound to perform better, we found out that a smaller model B3 beats a larger model B5 in accuracy and speed both. B3 has less than half the number parameters as B5. But we also found out that a very small model can fail too as with the case of B1 which performed significantly worse than B3 and B5 both.

We trained our model on Diabetic Retinopathy Detection 2015 [24] and APTOS 2019 [23] combined. It has the precision, recall, F1 score and quadratic weighted kappa of 81%, 80%, 80% and 91.25% respectively. In our work we used a slightly modified version of EfficientNet [25] and found out that using correct hyperparameters and adequate image processing, good results can be achieved. From the experimental results we conclude that the effectiveness and speed of our model is at par with other proposed models and can be deployed in clinical applications as an aid to ophthalmologists. We do not claim that it could replace an ophthalmologist but it surely can make DR detection faster.

Looking at the future prospects of Diabetic Retinopathy we can say that a smart phone device connected to a retinal image capturing camera will be efficient to diagnose DR in any area. This can help solve the value and time challenges of the conventional methods. Smartphone based automatic diagnosis of DR could be generally performed in two ways. The first and straightforward way could be a web-based diagnosis, and; the second could be a stand-alone independent smart phone software-based diagnosis. An independent smart phone application could enable patients living in rural and remote areas to urge an automatic diagnosis of DR easily. Another approach would be to have a cloud based diagnosis, which will enable users to always be updated and have the latest and greatest in DR research at there fingertips.

VI. REFERENCES

14. Mohammed Abbas Kadhim, Mohammed Hamzah Abed, Convolutional Neural Network for Satellite Image Classification, Intelligent Studies in Computational Intelligence DOI: 10.1007/978-3-030-14132-5_13