Diabetic Retinopathy Detection using Retinal Images and Deep Learning Model

Vani Ashok, Navneet Hosmane, Ganesh Mahagaonkar, Aditya Gudigar, Anvith P

Abstract: Diabetic Retinopathy (DR) is one of the serious problems caused by diabetes and a leading source of blindness in the working-age population of the advanced world. Detecting DR in the early stages is crucial since the disease generally shows few symptoms until it is too late to provide an effective cure. But detecting DR requires a skilled clinician to examine and assess digital color fundus images of the retina. By simplifying the detection process, severe damages to the eyes can be prevented. Many deep learning models particularly Convolutional Neural Networks (CNNs) have been tested in similar fields as well as in the detection of DR in early stages. In this paper, we propose an automatic model for detecting and suggesting different stages of DR. The work has been carried out on APTOS 2019 Blindness Detection Benchmark Dataset which contains around 3600 retinal images graded by clinicians for the severity of diabetic retinopathy on a range of 0 to 4. The proposed method uses ResNet50 (Residual Network that is 50 layers deep) CNN model along with pre-trained weights as the base neural network model. Due to its depth and better transfer learning capabilities, the proposed model with ResNet50 achieved 82% classification accuracy. The classification ability of the model was further analysed with Cohen Kappa score. The optimized validation Cohen Kappa score of 0.827 indicate that the proposed model didn't predict the outputs by chance.

Keywords: APTOS, Convolutional Neural Networks, Diabetic Retinopathy, ResNet50

I. INTRODUCTION

Diabetes affects around 400 million people worldwide, 70 million in India alone. Diabetic Retinopathy (DR) is an eye disease associated with long-standing diabetes.

Advancement to blindness can be slowed or averted if DR is detected in time, but this can be challenging as the disease rarely shows any symptoms in early stages until it is too late to give effective treatment. In the early stages DR is symptom-less, but that's when it is treatable. So, in ideal cases screening it early on helps a lot in preventing severe effects. During these early stages if we take retinal images of back of the eye there will be many bleeding spots on the retina. According to a survey, in India, there are about 11 ophthalmologists for every million people.

Keeping up with the sheer number of patients let alone focus on each individual is very difficult. The main problem is that not enough people undergo screening at the right time. And not every hospital has the tools and specialists required to detect DR. So, it is important to introduce methods that are easy to use and can be installed in even remote places.

These techniques, along with approaches grasping the potential of preventative or advancing medicine, could lead to better manage DR, including treatment at earlier stages and more specific fit of treatments based on individual patient variations. Detecting DR is a categorical problem, where given an eye image (Retinal) the system should predict the correct stage of DR the eye is in. DR progresses in 5 stages: 0 - No DR, 1 - Mild DR, 2 - Moderate DR, 3 - Severe DR, 4 - Proliferative DR. Each stage has a different characteristics and severity. Based on the retinal images taken for each of these stages, we need to extract features and train the model which predicts accurate severity.

II. RELATED WORK

Many researchers have put their time and effort in the early detection of DR. In early stages, many classical approaches of computer vision and machine learning were used to implement a suited solution to this problem. For example, E. V. Carrera et al. [1] recommended a computer-aided diagnosis based on digital processing of retinal images. SVM was used to classify retinopathy severity of each retinal image. This work produced a maximum sensitivity of 95% and a predictive capacity of 94%.

With increasing recognition of deep learning approaches, many techniques that apply CNNs to this problem showed up. By building a custom CNN architecture with 16 layers, R. Pires et al. [2] proposed a model which would predict if an image had a referable DR. Multi-image resolution and two-fold cross-validation were applied during training. The Convolutional Neural Network of the 512 x 512 image input was trained after initializing the weights by the trained CNN on a smaller image resolution. Kaggle dataset was used to train the CNN and Messidor-2 along with the DR2 dataset was used for testing the network.

This yielded an area under the Receiver Operating Characteristic (ROC) curve of 98.2% when testing with the Messidor-2 dataset.

One of the main reasons why deep neural networks are tough to train is because the gradient does not get well transferred to the input. This problem is called the vanishing gradient problem. We get another problem called degradation when deep neural networks start converging. To solve this problem many researchers used residual networks. For instance, M. T. Esfahan et al. [3] used a variant of CNN, which is ResNet34 in their research to classify DR images of the Kaggle dataset into DR or normal images. Resnet34 is one feasible CNN architecture pre-trained on ImageNet database. They reported an accuracy of 85% and a sensitivity of 86%.

III. METHODOLOGY

A. Dataset

The image set used to implement the DR detection model is taken from APTOS (Asia Pacific Tele-Ophthalmology Society) 2019 DR Detection competition in Kaggle [5]. Like any real-world dataset, there is noise in both the labels as well as images. Images may be underexposed, out of focus, or overexposed. The images were gathered from multiple clinics using a variety of cameras over an extended period, which will introduce further deviation. The dataset has around 3600 images with the severity of DR ranging from 0 to 4. The Distribution of classes in the dataset is shown in Fig. 1. In the dataset class 0 has most of the samples and class 4 has minimum samples compared to others.
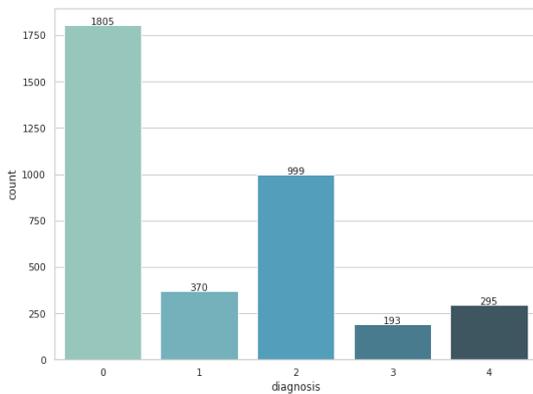

Fig. 1: Class distribution in dataset

Fig. 2 shows a sample image from the dataset. The image contains picture of an eye taken from a special camera.
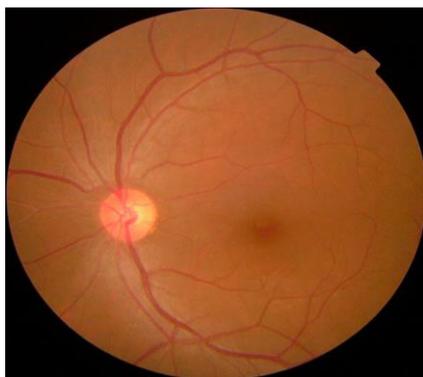

Fig. 2: Sample image from dataset

B. Image Preprocessing

Image Preprocessing requires each of the images to be resized to a standard size and apply filters to amplify feature extraction. A Gaussian blur (also known as Gaussian smoothing) is the result of masking an image by a Gaussian function (named after mathematician and scientist Carl Friedrich Gauss). It is typically used in image processing to reduce image noise and detail. The effect of this blurring technique on images is a smooth blur resembling that of viewing the image through a translucent screen [4]. By applying Gaussian blur, the leakage in blood vessels becomes more evident.

Along with Gaussian blur image resizing was used to make all images in the dataset of standard size. Also, images in training and validation sets were cropped to remove unwanted background. For this circle cropping technique was used, which crops the retinal images along the curvature of the eyes.

C. Data Augmentation

Data augmentation is one of the most used procedures to develop additional images from the dataset to make it generalize well on new data with rotation flips, cropping, padding, etc. Image Augmentation techniques used were – horizontal flip, optical distortion, grid distortion, piecewise affine transform, vertical flip, random rotation, random shift, random scale, random brightness and contrast, additive Gaussian blur, sharpening, embossing, and so on. Using these augmentation techniques, better insights about the images in the dataset were obtained.

D. Network Architecture

ResNet-50 is a CNN that is 50 layers deep. The network has a pre-trained version that is trained on more than a million images from the ImageNet database. ResNet architecture has a residual connection, from previous layers to the current one i.e., the summation of outputs from previous layers is given as input to the present layer. This results in faster learning of model and better accuracy. To address the degradation problem residual neural network fit explicitly the mapping, $F(x)+x$ - where x - input and $F(x)$ - mapping of the stack of non-linear layers (residual mapping). A basic block of such residual networks is shown in Fig. 3[6].
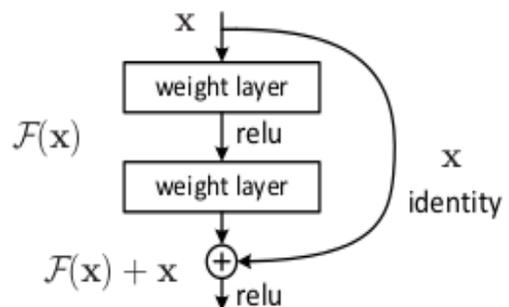

Fig. 3: Residual Learning: a basic block

Deep convolutional neural network models may take days or even weeks to train on very large datasets. A way around this problem is to reuse the model weights from pre-trained models, such as the ImageNet image recognition tasks. Efficient models can be integrated into a new model or used directly for different computer vision problems [7]. So, to train the model, pre-trained weights were used [8]. This increases the model performance and accuracy. Fig. 4 gives a general architecture of the ResNet50 model.
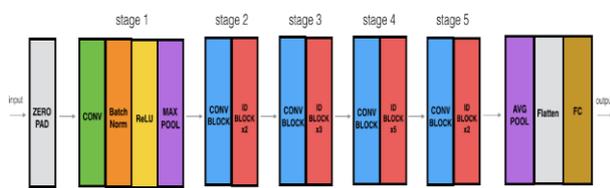

Fig. 4: ResNet50 architecture [9]

E. Training Process

Among the 3662 training images present in Kaggle's APTOS 2019 Blindness Detection dataset, 2930 were used in training the model and the remaining 732 were used for validating the model [5]. As we can see from Fig. 1 the training dataset is unbalanced, with most images belonging to class 0. But this is not an issue, since ResNet50 offers a solid learning model. Standardized images of size 512 X 512 were given as input to the model. Here ResNet50 architecture acts as a base model. To this base model, we connected the Global Average Pooling layer and an output Dense layer which consists of 5 nodes, each for different stages of DR present in the dataset. In the training process along with standard accuracy metrics, Cohen Kappa score was also used as an evaluation metric. Apart from the agreement that we observe from the Confusion matrix, Cohen Kappa narrates the agreement that occurs by chance.

IV. RESULT AND DISCUSSION

A. Data Augmentation Results

As shown in Fig. 5, images are rotated and zoomed to get better insights from existing training images. Images present in the training image set are duplicated by applying zoom of range 1/1.22, rotation range of 180 ^{0}, and other operations which help bring out more insights.
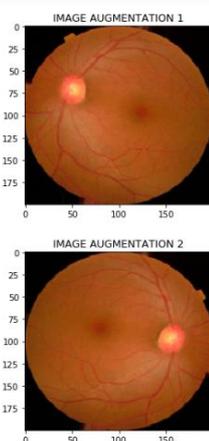

Fig. 5: Image augmentation results

B. Model Evaluation Metrics

- **Accuracy:** Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations [10].
- **Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate [10].
- **Recall:** Recall is the ratio of correctly predicted positive observations to all the observations in actual class [10].
- **F1 Score:** F1 Score is the weighted average of Precision and Recall [10].
- **Cohen Kappa Score:** To validate that the results obtained in the above evaluation metrics were not by chance, the Cohen Kappa score was also calculated both for training and validation steps. We obtained a training Cohen Kappa score of 0.957, validation Cohen Kappa score of 0.812 and a Complete set Cohen Kappa score of 0.93.

For different classes in the proposed dataset, model performance on the validation dataset is as depicted in Table-I. The metrics for classes '3' and '4' are very less since their support i.e., The number of samples are less. Considering this, the scores obtained are satisfactory. The overall accuracy obtained is 82%, macro avg, which will aggregate the contributions of all classes to compute the average metric is 65% and weighted avg, which is averaging the support-weighted mean per label is found to be 83% on average.

Table- I: Performance of ResNet50

Classes	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	352
1	0.65	0.64	0.65	73
2	0.83	0.73	0.78	237
3	0.43	0.39	0.37	41
4	0.27	0.59	0.37	29
Macro Avg	0.64	0.67	0.64	732
Weighted Avg	0.84	0.82	0.83	732
Accuracy		0.82		732

C. Model Accuracy Plot

Model accuracy plots help to visualize and review the performance of deep learning models over time during training by plotting the Number of Epochs Vs. Accuracy. It represents how the model improves with the number of epochs. As shown in Fig. 6, the model overfits if the number of epochs is more than 20. So, it was trained for 20 epochs and maximum validation accuracy of 82% was achieved from the 12th epoch.
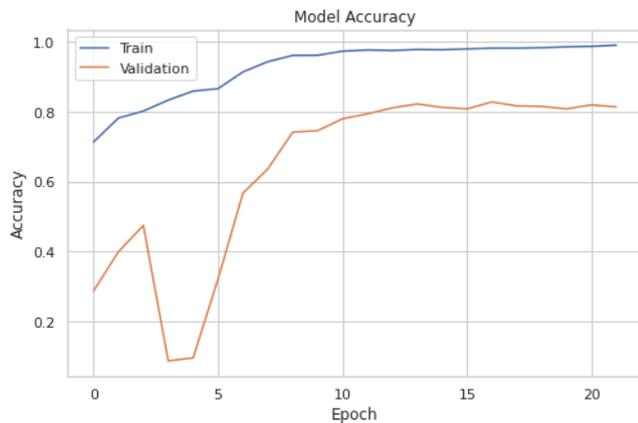
Fig. 6: Model accuracy plot

D. Confusion Matrix

A confusion matrix is used to express the performance of a classification model on validation data. In the confusion matrix shown in Fig. 7, primary diagonal elements indicate the True Positives which is 347 for class 0, 47 for class 1, 174 for class 2, and so on. The remaining elements of the matrix indicate True and False negatives for the respective classes. The mispredictions are due to the unbalanced dataset used to train the model.
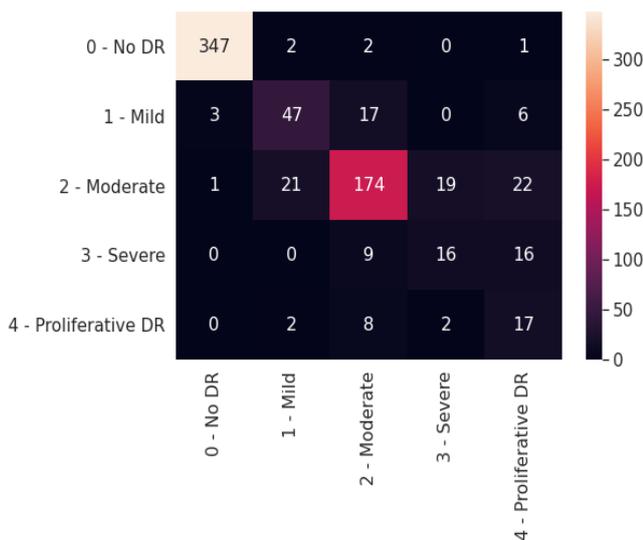

Fig. 7: Confusion matrix for validation data

E. Testing Results

With the test dataset, predicted results of the model is shown in Fig. 8. Most of the images belonged to class 2 - Moderate and the second most frequent class belonged to class 0 – No DR. This is because classes 0 and 2 have most samples in the training dataset. But notable point is that the model predicted minority classes like 3 and 4 in some of the cases. This shows that the model even though given with unbalanced data has learned features from minority classes too.
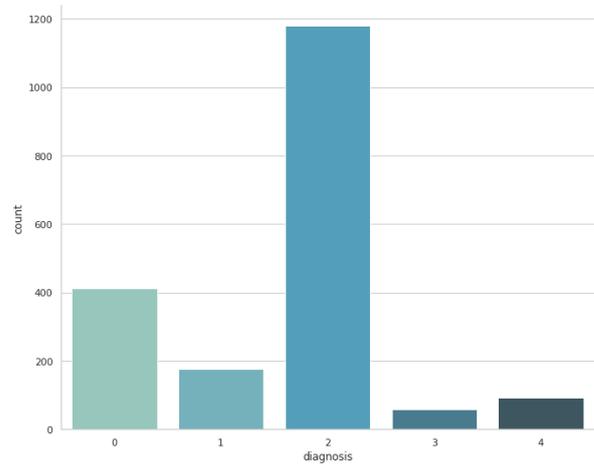

Fig. 8: Results for test dataset

V. CONCLUSION

Automated DR detection systems remarkably shorten the time required to determine diagnoses, saving effort, costs and resulting in the timely treatment of patients. For detecting DR at an early stage, a screening system for DR detection plays an important role. Based on the type of lesions that appear on the retinal images obtained from screening, the stage of DR is detected.

In the proposed work, it is proved that even automated systems can go par with manual methods of detecting DR. The proposed pre-trained model was able to classify the retinal images into 5 classes based on severity of diabetic retinopathy (ranging from 0 to 4) with 96% and 82% train and validation accuracy, respectively. The performance of the model was further analysed with different evaluation metrics such as precision, recall, F1-score. These results indicate that the classification behaviour was extremely well for some classes and satisfactory for remaining classes considering the fact that the augmented dataset used was an unbalanced one. To better explain the classification accuracy, the Cohen Kappa score was also calculated. The optimized validation Cohen Kappa score of 0.827 indicate that the proposed model didn't predict the outputs by chance. But with any Deep Learning model applied in the medical field, size and quality of the dataset is an issue. To give accurate results just as manual methods of detecting DR, we need to further improve our model.

REFERENCES

1. E. V. Carrera, A. Gonzalez, and R. Carrera, *"Automated detection of diabetic retinopathy using SVM,"* 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), 2017, pp. 1-4.
2. R. Pires, S. Avila, J. Wainer, E. Valle, M.D. Abramoff, A. Rocha, *"A data-driven approach to referable diabetic retinopathy detection Artif Intell Med"*, 96 (2019), pp. 93-106.
3. M.T. Esfahani, M. Ghaderi, R. Kafiyeh, *"Classification of diabetic and normal fundus images using new deep learning method Leonardo Electron J Pract Technol"*, 17 (32) (2018), pp. 233-248.
4. https://en.wikipedia.org/wiki/Gaussian_blur
5. https://www.kaggle.com/c/aptos2019-blindness-detection/data
6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. *"Deep Residual Learning for Image Recognition"*. Available: https://arxiv.org/pdf/1512.03385.pdf

7. https://machinelearningmastery.com/how-to-use-transfer-learning-when-developing-convolutional-neural-network-models/
8. https://github.com/fchollet/deep-learning-models/releases/tag/v0.1
9. https://towardsdatascience.com/residual-networks-resnets-cb474c7c834a
10. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

AUTHORS PROFILE

Vani Ashok, obtained her master's degree in computer engineering in the year 2003. She is pursuing research in the area of image processing and artificial intelligence. She is working as Assistant Professor in the Department of Computer Science and Engineering at JSS Science and Technology University, Mysuru, Karnataka, India. She has presented papers in international conferences and published papers in international journals.

Navneet Hosmane, B.E., Department of Computer Science and Engineering, 4th year student at JSS Science and Technology University, Mysuru.

Ganesh Mahagaonkar, B.E., Department of Computer Science and Engineering, 4th year student at JSS Science and Technology University, Mysuru.

Aditya Gudigar, B.E., Department of Computer Science and Engineering, 4th year student at JSS Science and Technology University, Mysuru.

Anvith P, B.E., Department of Computer Science and Engineering, 4th year student at JSS Science and Technology University, Mysuru.