Deep Learning Approach to Diabetic Retinopathy Detection

Deepak Gaur (222IT008)
Information Technology
National Institute of Technology, Karnataka, Surathkal, India 575025
deepakgaur.222it008@nitk.edu.in

Natasha Jain (222IT023)
Information Technology
National Institute of Technology, Karnataka, Surathkal, India 575025
natashajain.222it023@nitk.edu.in

Dr. Dinesh Niak
Information Technology
National Institute of Technology, Karnataka, Surathkal, India 575025
din_nk@nitk.edu.in

***Abstract* - If unchecked, diabetic retinopathy is one of the most dangerous consequences of diabetes, since it may cause a person to become blind for life. The importance of early diagnosis to the overall outcome of therapy makes it one of the primary obstacles to overcome. Unfortunately, the precise diagnosis of the diabetic retinopathy stage is notoriously difficult and needs the skilled interpretation of fundus pictures by a person. Simplifying the stage of detection is essential and has the potential to benefit millions of individuals. There have been several successful applications of convolutional neural networks (CNN), including for the diagnosis of diabetic retinopathy itself. However, the efficacy of these approaches is hindered by the prohibitively high cost of acquiring large labelled datasets as well as variability across various physicians. In this research, we present an automated technique for stage identification of diabetic retinopathy based on deep learning that utilises a single image of the fundus of a human subject. In addition, we present the multistage method to transfer learning, which takes use of datasets that are comparable but have been labelled differently. The presented method has a sensitivity and specificity of 0.99, making it suitable for use as a screening method for the early detection of diabetic retinopathy. On the APTOS 2019 Blindness Detection Dataset, it was ranked 54 out of 2943 competing methods, with a quadratic weighted kappa score of 0.925466, according to the methodology's position (13000 images).**

Keywords: Deep learning, diabetic retinopathy, deep convolutional neural network, multi-target learning, ordinal regression, classification, SHAP, Kaggle, APTOS.

I. INTRODUCTION

One of the most serious and potentially blinding effects of diabetes is called diabetic retinopathy (DR). DR is characterised by damage to the retina, which ultimately leads to vision loss. It causes the blood vessels inside the retinal tissue to become damaged, which results in the vessels leaking fluid and resulting distortion of vision. According to the data from the United States, the United Kingdom, and Singapore (NCHS, 2019; NCBI, 2018; SNEC, 2019), DR is one of the most common conditions. It is in the same category as diseases that might lead to blindness, such as cataracts and glaucoma.

DR progresses with four stages:

- Mild non-proliferative retinopathy is the first stage of the disease and is the only stage in which microaneurysms may develop.
- Moderate non-proliferative retinopathy is a stage of the illness that may be defined as the blood vessels losing their capacity to carry blood owing to the vessels being distorted and swollen as the disease progresses.
- Because of the increasing blockage of more blood vessels, severe non-proliferative retinopathy causes a reduction in the blood flow to the retina, which in turn sends a signal to the retina to begin the process of developing new blood vessels;
- The advanced stage of diabetic retinopathy is called proliferative diabetic retinopathy. This condition occurs when the growth features that are secreted by the retina activate the proliferation of new blood vessels. These new blood vessels grow along the inside covering of the retina in some vitreous gel, which fills the eye.

Because each stage has its own traits and special qualities, it is possible for medical professionals to fail to take into consideration some of those aspects, leading to an inaccurate diagnosis. The conclusion that can be drawn from this is that an automated method of DR detection should be developed.

At least 56% of new instances of this condition might be reduced with therapy that is both appropriate and timely as well as close monitoring of the eyes (Rohan T, 1989). However, since the early stage of this illness does not provide any warning symptoms, it is very difficult to identify the condition in its first stages. In addition, even very experienced medical professionals arXiv:2003.02261v1 [cs.LG] 3 Mar 2020 could not manually analyse and assess the stage using diagnostic photographs of a patient's fundus, according to study conducted by Google (Krause et al., 2017), which is seen in figure 1. On the other hand, medical professionals will almost always reach the same conclusion when lesions are present. In addition, current methods of diagnosis are highly inefficient owing to the length of time they take and the number of ophthalmologists who are involved in the process of finding a solution to the patient's condition. These different points of view lead to incorrect diagnoses and shaky ground truth for the automated solutions that were developed to aid in the research stage.

II. PROBLEM STATEMENT

A. Dataset

The picture data that was used for this study came from a variety of different sources. When we were pretraining our CNNs, we made use of an available dataset that was provided by the Kaggle Diabetic Retinopathy Detection Challenge 2015 (EyePACs, 2015). This dataset is the most extensive one that is open to the public. It is made up of 35126 photos of the fundus of the left and right eyes of American individuals, each of which is labelled with a stage of diabetic retinopathy: early, moderate, and severe.

- No diabetic retinopathy (label 0)
- Mild diabetic retinopathy (label 1)
- Moderate diabetic retinopathy (label 2)
- Severe diabetic retinopathy (label 3)

In addition, we used other smaller datasets: Indian Diabetic Retinopathy Image Dataset (IDRiD) (Sahasrabuddhe and Meriaudeau, 2018), from which we used 413 photographs of the fundus, and MESSIDOR (Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology) (Decencire et al., 2014) dataset, from which we used 1200 fundus photographs. As the original MESSIDOR dataset has different grading from other datasets, we used the version that was relabeled to standard grading by a panel of ophthalmologists (Google Brain, 2018).
As the evaluation was performed on Kaggle APTOS 2019 Blindness Detection (APTOS2019) dataset (APTOS, 2019), we had access only to the training part of it. The full dataset consists of 18590 fundus photographs, which are divided into 3662 training, 1928 validation, and 13000 testing images by organizers of Kaggle competition. All datasets have similar distributions of classes; distribution for APTOS2019 is shown in Figure 2.

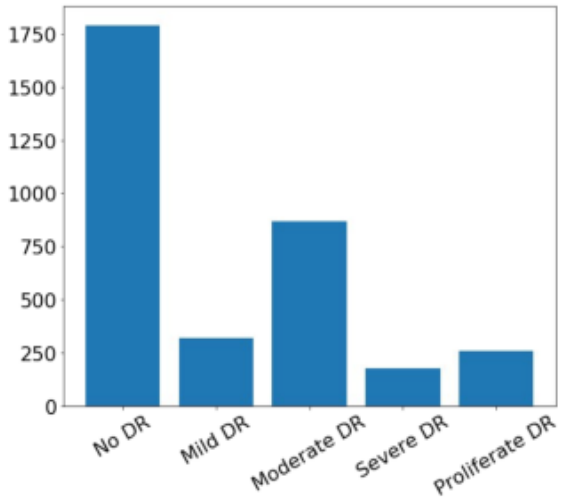

Figure 2: Classes distribution in APTOS2019 dataset.

As different datasets have a similar distribution, we considered it as a fundamental property of this type of data. We did no modifications to the dataset distribution (undersampling, oversampling, etc.).
The smallest native size among all of the datasets is 640x480. Sample image from APTOS2019 is shown in Figure 3.

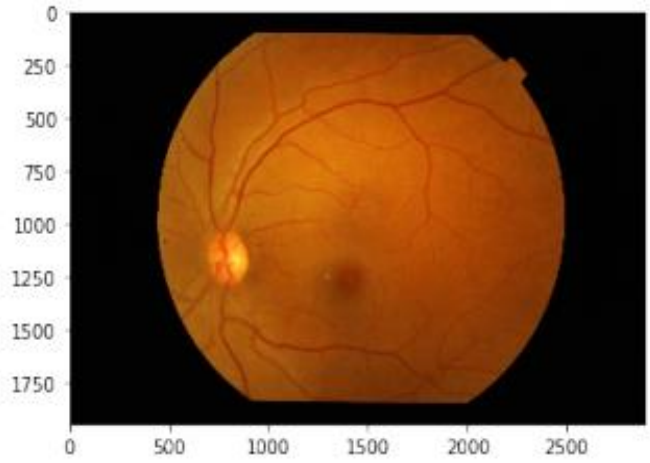

Figure 3: Sample of fundus photo from the dataset.

B. Evaluation Metric

As the primary statistic that we employed in this investigation, the quadratic weighted Cohen's kappa score was used. A Kappa score indicates the degree to which two evaluations are consistent with one another. The quadratic weighted kappa is a statistic that compares the scores that were given by human raters to the scores that were anticipated based on those ratings. This statistic ranges from -1, which indicates that raters are completely in agreement, to 1. (complete agreement between raters). The following is one definition of k:

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} o_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} e_{ij}}, \qquad (1)$$

If k is the total number of categories, oi j and ei j are the corresponding numbers of elements in the observed and predicted matrices, and k is the total number of categories. The formula for calculating wi j is as follows:

$$w_{ij} = \frac{(i-j)^2}{(k-1)^2}, \qquad (2)$$

Because of the features of Cohen's Kappa, researchers have to interpret this ratio very carefully. For example, if we take two pairs of raters with the same percentage of an agreement but different proportions of ratings, we should be aware that this will have a significant impact on the Kapparatio.

Another issue is the large amount of codes, which causes Kappa to increase in proportion to its size. Additionally, the Kappa statistic may be low despite the fact that there are high levels of agreement and that individual evaluations are correct. Because of all that was just said, Kappa is a ratio that is very difficult to examine.

The fact that we do not have access to the labels of the validation and testdatasets is the primary reason why we make use of the Kappa ratio. The Kappa value for these datasets may be acquired by submitting our model and the runner's code to the validation system that is available on the Kaggle website. In addition, we are unable to have direct access to the photos that are included in the test dataset.

In addition to the Kappa score, we compute the macro F1- score, accuracy, sensitivity, and specificity on a holdoutdataset consisting of 736 pictures drawn from the APTOS2019 training data.

III. METHOD

The challenge of detecting diabetic retinopathy may be looked at from a number of different perspectives, including that of a classification problem, a regression problem, and an ordinal regression problem (Ananth and Kleinbaum, 1997). This is a possibility due to the progression of the illness, which occurs in phases.

A. Preprocessing

Both the training of the model and its validation were carried out using preprocessed copies of the raw pictures. The first step in the preprocessing was cropping the picture, then the next step was resizing it.

Because of the manner that the data for APTOS2019 was obtained, there are false correlations between the stage of the illness and a number of picture meta-features, such as the resolution, croptype, zoom level, or overall brightness of the image. Figure 4 provides a representation of the correlation matrix.

We employed a large number of augmentations in order to prevent CNN from being overfit to these characteristics and to lessen the correlations that existed between the picture content and its meta-features. In addition, given that we cannot access the test dataset during either the competition or in real life, we have made the decision to provide models with as much data variation as is practically feasible.

B. Data Augmentation

We made advantage of online augmentations, and the training picture was modified with at least one of these modifications before being fed into the CNN. We made use of the following augmentations from the Albumentations (A. Buslaev and Kalinin, 2018) library: optical distortion, grid distortion, piecewise affine transform, horizontal flip, vertical flip, random rotation, random shift, random scale, a shift of RGB values, random brightness and contrast, additive Gaussian noise, blur, sharpening, embossing, randomgamma, and cutout (Devries and Taylor, 2017).

C. Network Architecture

Our objective is to properly categorise each fundus image. Conventional deep CNN design, which contains a feature extractor and a smaller decoder for a given job, is what we use to construct our neural networks (head).

However, training the encoder from scratch is challenging, particularly considering the limited quantity of training data available. Therefore, as initialization for the encoder, we employ an Imagenet-pretrained CNN (Iglovikov and Shvets,2018).

In order to diagnose diabetic retinopathy, we advocate for the multi-task learning methodology. We use a total of three decoders. Each is taught to tackle its own job using the characteristics obtained from the CNN backbone:

- classification head,
- regression head,
- ordinal regression head.

In this situation, the classification head generates a one-hot encoded vector, in which the existence of each step is denoted by the number 1. After producing a real number in the range [0;4:5], the regression head generates a number that is then rounded to produce an integer that reflects the illness stage. In order to calculate the head of the ordinal regression, we apply the method outlined in (Cheng, 2007). To summarise, if the data point is classified as belonging to category k, then it is also classified as belonging to all of the categories that range from 0 to k minus 1. As a result, the goal of this head is to make predictions across all categories up to the objective. The final prediction is achieved by fitting a linear regression model on the outputs of threeheads. This results in the final prediction. The structure of a neural network is seen in Figure 5. In order to cut down on the total amount of time spent training, we train all heads and the feature extractor together. Before moving on to the post-training step, we do not defrost the linear regression model.

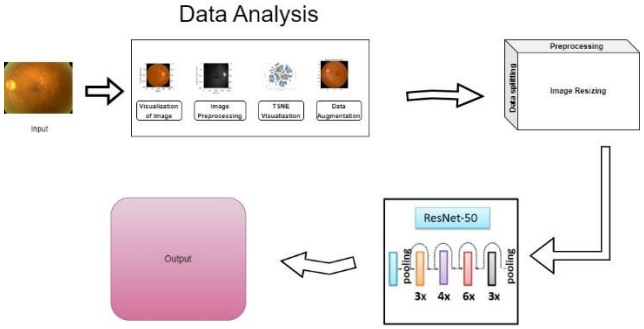

Fig:4 Components of the project

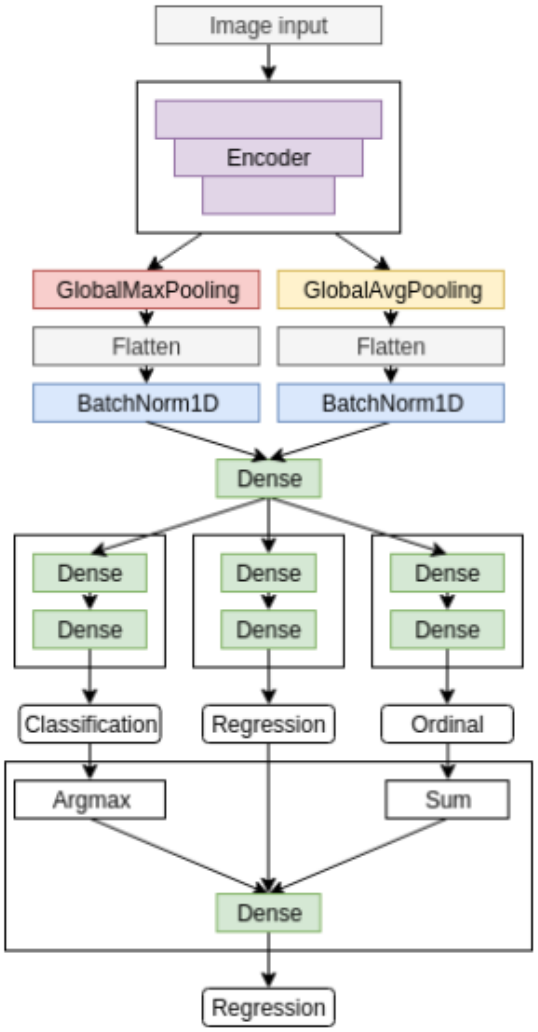

Figure 5: Three-head CNN structure.

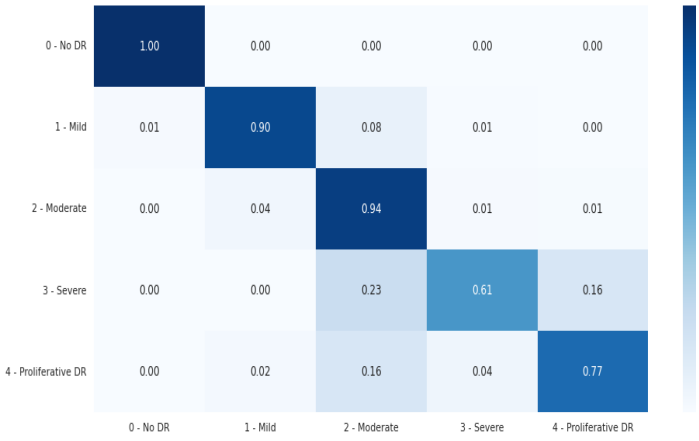

Fig :6 Confusion matrix

IV. RESULTS

After the data set results found were as follows

1. The precision for class 0 was 1.00, recall value was 1.00 f1-score was 1.00 and support was 1805
2. The precision for class 1 was 0.87, recall value was 0.90 f1-score was 0.89 and support was 370
3. The precision for class 2 was 0.94, recall value was 0.94 f1-score was 0.91 and support was 999
4. The precision for class 3 was 0.81, recall value was 0.61 f1-score was 0.869 and support was 193
5. The precision for class 4 was 0.84, recall value was 0.77 f1-score was 0.81 and support was 295

The accuracy of F1 score was 0.93 with a support of 3662

The macro avg was 0.88,0.84, 0.86 , 3662 of precision , recall, f1 score and support and the weighted avg was 0.93 , 0.93 , 0.93 and 3662 for precision , recall, f1 score and support

The ensembles received a QWK score of 0.959

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1805
1	0.87	0.90	0.89	370
2	0.88	0.94	0.91	999
3	0.81	0.61	0.69	193
4	0.84	0.77	0.81	295
accuracy			0.93	3662
macro avg	0.88	0.84	0.86	3662
weighted avg	0.93	0.93	0.93	3662

Table 1: Results

```
                                                                    ]

dense (Dense)                (None, 2048)              4196352     ['dropout[0][0]']

dropout_1 (Dropout)          (None, 2048)              0           ['dense[0][0]']

final_output (Dense)         (None, 5)                 10245       ['dropout_1[0][0]']

==================================================================================================
Total params: 27,794,309
Trainable params: 27,741,189
Non-trainable params: 53,120
__________________________________________________________________________________________________
```

Table 2: Summary of the model

V. INTERPRETATION

When it comes to medical applications, having the ability to comprehend the predictions that models make is essential. Although a strong performance on the validation dataset may be used as a criterion for choosing the best-trained model for production, this alone is not adequate for making use of this model in real-world settings.

It is feasible to depict characteristics that help to the evaluation of the illness stage by making use of SHAP (Shapley Additive exPlanations), which was developed by Lundberg and Lee (2017). SHAP brings together a number of different ways that have been used in the past and is the only method that is able to be both consistent and accurate while adding features locally.

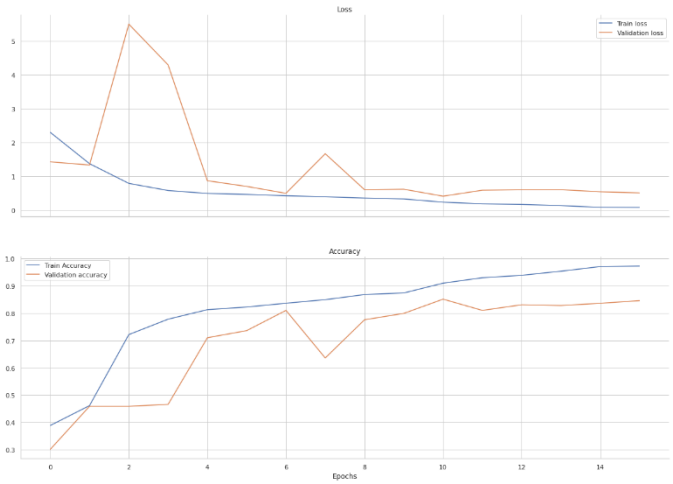

Fig:7 Loss vs Validation graph

By using SHAP, one can guarantee that the model will acquire important features during the training phase and will employ the proper features when it comes to making inferences. In addition, the visualisation of important traits may be of assistance to the physician in unclear situations by directing their attention to areas of interest in which features are most readily apparent.

VI. CONCLUSION

In this project, we introduced the multistage transfer learning strategy as well as an automated method for detecting the stage of diabetic retinopathy by singlephotography of the human fundus. Both of these methods may be found in the accompanying table. For the purpose of developing our final solution, we made use of transferlearning and an ensemble of three different CNN architectures (EfficientNet-B4, EfficientNet-B5, and SE- ResNeXt50). The results of the experiments reveal that the suggested strategy is capable of achieving high and steady outcomes even when the metric being used is unstable. This method's primary benefit is that it improves generalisation while simultaneously reducing variance. This is accomplished by making use of an ensemble of networks that have been pretrained on a large dataset and fine-tuned using the target dataset. In further work, it may be possible to enhance this technique by computing SHAP not just for a specific network but also for the whole ensemble, and by optimising hyperparameters in a manner that is more precise. In addition, we are able to conduct studies with pretrained encoders on various tasks associated with eye conditions. Additionally, it is feasible to examine meta-learning (Nichol et al., 2018) using these models; however, it should be understood that this will need for independent in-depth research.

VII. ACKNOWLEDGMENT

We would like to thanks and express our gratitude to Prof. Dr Dinesh Naik for their valuable feedback that greatly helped us for the better implementation of our work.

VIII. REFERENCES

[1] A. Buslaev, A. Parinov, E. K. V. I. I. and Kalinin, A. A. (2018). Albumentations: fast and flexible image augmentations. *ArXiv e-prints* H. Simpson, *Dumb Robots*, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.

[2] Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology*, 26(6):1323–1333..

[3] Christopher E.Hann, J. Geoffrey Chase, J. A. R. D. H. G. M. S. (2009). Diabetic retinopathy screening using computer vision J.-G. Lu, "Title of paper with only the first word capitalized," *J. Name Stand. Abbrev.*, in press.

[4] Yung-Hui Li, Nai-Ning Yeh, S.-J. C. and Chung, Y.-C. (2019). Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network. M. Young, *The Technical Writer's Handbook*, Mill Valley, CA: University Science, 1989.