Transfer Learning approach for grading of Diabetic Retinopathy

T Aswathi¹, T R Swapna² and S Padmavathi²

¹ Student Research Scholar, Department of Computer Science Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
² Assistant Professor, Department of Computer Science Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.

aswathi9434@gmail.com

Abstract. There has been a wide interest in applying Deep Learning (DL) algorithms for automated binary and multi class classification of colour fundus images affected with Diabetic Retinopathy (DR). These algorithms have shown high sensitivity and specificity for detecting DR in non-clinical setup. Transfer learning has been successfully tested in many medical imaging applications like skin cancer detection, pulmonary nodule detection, Alzheimer’s disease etc. This paper experiments with the different DL architectures such as VGG19, InceptionV3, ResNet50, MobileNet and NASNet for automated DR classification (binary and multi class) on Messidor dataset. The dataset is publicly available, and comprises of 1200 retinal fundus images. The images belong to four different classes of DR namely, normal (class 0), mild (class 1), moderate (class 2) and severe (class 3), graded based on the severity level of DR. In our experiment, we have enhanced the quality of input images by applying algorithms like CLAHE (Contrast Limited Adaptive Histogram Equalisation) algorithm and Powerlaw transformation as pre-processing techniques, which work on the small image patches with high accuracy, contrast limiting and image sharpening. Hyper parameter tuning on pretrained InceptionV3 architecture, resulted in enhancing the accuracy of the model. Both binary and multi class results were analysed considering inter class (one class with another class) accuracies. We achieved an accuracy of 78% between class 0 and class 1, the accuracy between class 0 and class 2 further reduced to 69%, while class 1 and class 2 showed an accuracy of 61%. Moreover, the interclass class accuracy between class 1 and class 3 was 62%, class 2 and class 3 further reduced to 49%. The accuracy further diminished between class 0 and class 3 to 32%. These experiments suggest that the pretrained models provided better results in classifying normal and mild, but they were not that much efficient in classifying moderate-severe and normal-severe binary classifications.

1. Introduction

According to International Diabetes Federation (IDF), there are 77 million diabetic patients in India and prevalence of diabetes in adults is 8.9% in the year 2020 [1]. When diabetes affects retinal blood vessels, it results in DR, which is the main cause for blindness. Consultation with ophthalmologist and timely detection can reduce complications of blindness and the disease. Screening for poorly controlled diabetes should be done at least every year, but hardly anyone does it. Reasons for that could vary from lack of access to an ophthalmologist, limitations of screening procedure, unaffordable consultation cost and so on. These constraints in screening makes computer-based assessment of
retinal images that employs Artificial Intelligence (AI) based grading systems, extremely relevant. These systems help to determine the severity of the illness, and pursuing of treatment, which would have been very expensive in the ordinary method. Several studies have been made to analyse the efficacy of AI grading systems on detecting and grading of different stages of DR [3], [7] and [10].

There are mainly five stages for detection and grading of DR; they are normal, mild, moderate, severe and end stage [6] and [2]. The normal, mild, moderate and severe stages are together called Non-Proliferative Diabetic Retinopathy (NPDR) and the end stage is called Proliferative Diabetic Retinopathy (PDR). Microaneurysms, exudates, inflammation of the blood vessels in the eye are few of the primary symptoms of DR. Microaneurysms are the earliest clinically observable changes in DR detection. They are capillary dilation that are localised and typically saccular, which do not affect vision and look like small red dots that often appears in cluster or isolation. Exudates are the accumulation of lipids in retina, which are mainly classified into hard exudates and soft exudates. PDR is the most severe case of DR, in which new aberrant blood vessels are formed in various part of retina, termed as neovascularization, which leads to vision loss and irreversible blindness. Therefore, detection and grading of DR have a key role to play in diabetic patient's eye screening and treatment.

In computer vision applications, Convolutional Neural Networks (CNN) are mainly used for image analysis, classification and segmentation [11], [15] and [18]. Convolution layer, max pooling layer and fully connected layer are the commonly used layers in CNN architecture. Initial layers of the CNN capture more generic features like edges information while, end layers learn features that specific to the input dataset. So, in DR detection edge Information of the retinal images are captured in the initial layers while, features discriminating various class are learned in final layers. In CNN input matrix and filter matrix are multiplied together to get the convolutional output. Stride and padding are the parameters of CNN filters. The number of shifts made by filter matrix during the convolution process is called stride. Stride can be done both horizontally and vertically. If image size is not equal to the filter size, padding operations are performed. Pooling layer is used to bring down the size of the image matrix. Fully connected layer serves as a classifier in the final learning phase. The activation function will classify the input image into a specific class image. Input of the fully connected layer is a vector, converted from image features. Over fitting is reduced using dropouts, which randomly eliminate some layers of the CNN. Softmax layer is the final stage of fully connected convolution network and it generates the probability value of each class. The highest probability value indicates the class of the input object. In DR detection it identifies the grade of the specific retinal fundus image belongs to.

In case of DR, many DL methods have been broadly used for automatic classification tasks. The relevant features are retrieved using convolutional layers and final classification is done using fully connected layers with these features [2]. Gradient based optimization algorithms are used during training cases in order to change the model parameters. Ultimately softmax layer predicts the probabilities of input object belonging to the classes predefined in the model. The classification result is compared against the actual values present in the labeled dataset and calculates the accuracy [14].

Following sections discusses the literature surveys carried out for the study, the methodologies adopted for dataset selection that are CNN architectures, data augmentation and preprocessing, training and hyperparameter tuning, and the experiments carried out. The study then inspects the results of the experiments. It is followed by conclusion and future scope of the work.

2. Literature Survey

Various quantification approaches for DR detection without using DL are discussed in the study [8]. The results made more avidity in analysing the performance of DL based algorithm for DR detection tasks. The study [2] evaluates the efficiency of DL algorithm for automated identification of DR. It is found that DL algorithms considerably have high sensitivity and specificity for DR detection tasks. 128,175 retinal images were collected from EYEPACS dataset and three other eye hospitals in India. None, mild, moderate and severe were the grades chosen for annotating the dataset. The pretrained InceptionV3 [19], is used to train the model to make multiple binary predictions. The model could achieve high specificity and sensitivity on EYEPACS dataset and Messidor-2 dataset. An analysis [10]
of the detection of DR using DL, demonstrated that the performance of the DL model decreases with increase in number of classes. The analysis has used different CNN architectures to determine the best CNN for binary classification of DR images. Dataset chosen for the analysis was Kaggle and Messidor with CLAHE as preprocessing technique. Various other image enhancing techniques are discussed in studies [9], [12] and [13]. For the study [10] pretrained GoogleNet model was used for training, and hyperparameter tuning was done with batchnorm, L2 regularization, dropout, learning rate etc. The model was able to achieve 74.50% of test accuracy for binary classification, while 68%, 51% were the accuracy values for 3-ary and 4-ary classifications respectively. InceptionV3 architecture pretrained on ImageNet dataset with 41,122 images, was used in the study [7]. The model was capable of attaining a sensitivity of 89.6%, specificity of 97% and Area Under Curve (AUC) of 98%. The study [5] explains the relevance of deeper level fine-tuning for medical images. The work explains the importance of using fine-tuned pretrained model instead of training from scratch. The findings state that the deeper fine-tuning works better than shallow fine-tuning, compared to the results of CNN trained from scratch. [4] and [6] shows the reviews made in automatic detection of DR grading. The baseline considered in the study [4] were datasets, preprocessing techniques, ML based approach, DL based approach and performance measures, while for the study [6] were retinal image modalities, retinal anatomical structure, disease manifests, DL methods and segmentation methods. The study [3] analyses the ability of AI based system for DR detection implemented in the primary care practice. From 193 patient samples, the model could identify only 2 patients with disease and wrongly classified 15 patients as false positive. The small sample size and lack of generalizability were the drawbacks of the system. It is observed that deploying DL models for DR detection in clinical practice requires further assessment and validation. Our study consolidates the findings from all approaches and tries to find out where the DL systems are failing in the detection of the DR severity grades.

3. Methodology
The workflow diagram of the study is depicted in figure 1. The input images were initially split into training, validation and testing set with a ratio 60:20:20 respectively. All the images were then undergone preprocessing using CLAHE and powerlaw transformation then fed to pre-trained CNN architecture trained on ImageNet dataset. Hyper parameter tuning was performed, followed by training. The probabilities for belonging to each class were calculated as the classification result. Details of the workflow are described in below subsections.

![Figure 1. Methodology.](image)

3.1. Dataset
The Messidor dataset used for this study was created by three ophthalmologic departments with a color video 3CCD camera mounted on a Topcon TRC NW6 non-mydriatic retinograph and a 45-degree field of view. 8 bits per color plane were used to capture the images at 2304*1536,1440*960, or 2240*1488 pixels. 800 images were obtained with pupil dilation and 400 without pupil dilation.
The dataset was divided into four DR severity levels based on number of microaneurysms, hemorrhages, and neovascularization. Class zero contains 546 images and class one, two and three contains 153, 247 and 254 images respectively. The table 1 shows the criteria chosen for dividing the dataset into four severity levels. Details of the workflow are described in below subsections.

| Grades | Description |
|--------|-------------|
| 0      | ($\mu$A = 0) and (H = 0) |
| 1      | ($0 < \mu$A $\leq 5$) and (H = 0) |
| 2      | (($5 <\mu$A $< 15$) or ($0 < H < 5$)) and (NV = 0) |
| 3      | ($\mu$A $\geq 15$) or (H $\geq 5$) or (NV = 1) |

$\mu$A: Microaneurysms count  
H: Haemorrhages count  
NV = 1: Presence of neovascularization  
NV = 0: Absences of neovascularization

### 3.2. CNN Architectures
The CNN architectures used for the binary classification were InceptionV3, ResNet50 and VGG19 while the multi class classification was performed on two additional architectures-MobileNet and NASNet. Pretrained InceptionV3 has been chosen as the baseline architecture for further preprocessing tasks, hyperparameter tuning and other experiments. The comprehensiveness of the InceptionV3 architecture was instrumental in generating better results compared to other architectures for classification tasks. It is found that architecture performed comparatively better than other architectures in DR detection tasks. The InceptionV3 architecture has trained over 1 million images of ImageNet dataset. The architecture consists of convolutional layers, max pooling layers, global average pooling layers and also an Inception module. The network optimization and over-fitting problems are handled using dropouts. The whole architecture consists of 8 inception modules and 2 auxiliary classification modules, which are used for regularisation purpose. Depth and width of the architecture are balanced well so that information flow is maximum throughout the network.

### 3.3. Data augmentation and Pre-processing
Horizontal flipping, vertical flipping, rescaling etc., were the common data augmentation operations performed in the 1200 input retinal fundus images. CLAHE algorithm was used as the initial preprocessing method. Later, Powerlaw transformation was applied to enhance the quality of the images. As the initial step, the RGB (Red, Green, Blue) images were transformed to HIS (Hue, Intensity, Saturation) image. Luminance part of the image i.e., I component underwent CLAHE processing keeping the other two components as constant. In order to connect wide range of input values to narrow range of output values, the powerlaw transformation was performed on the I component. H and S components along with modified luminance component I were transferred back to RGB model to obtain the final preprocessed output.

### 3.4. Training and Hyper parameter tuning
For multi class classification, the fundus images were trained on VGG19, InceptionV3, ResNet, MobileNet and NASNet. All the architectures were pretrained on ImageNet dataset which consists of 1,000 classes. The images were resized into the input size of each architecture. Transfer learning was performed using all the pretrained architectures. A drop out of 0.5 was applied and dense layer was added with number of classes as 4. The results of the multi class classification is explained in the section below. The binary class classification was performed with best three architectures in the multi class classification that were, VGG19, InceptionV3, and ResNet. In binary class classification task,
class zero and class one of the datasets were combined together as one class, and class two and class three of the dataset were combined together as another class. Training procedure was the same as multi class classification except the dense layer with number of classes as 2 was added on top of the pretrained model. Results of binary class classification with each of the architectures using different optimizers are shown in the below sections. Learning rate, batch size, momentum, activation functions etc., were the parameters taken for hyperparameter tuning. The values of the parameters were chosen in such a way that they yield the best results for the pretrained InceptionV3 architecture from previous studies [3], [4], and [7]. Results of the hyperparameter tuning are covered in section 4.

3.5. Experiments
In order to analyse the contribution of each class for the final classification process, the performance of each class was measured against other classes. The experiment was carried out in the same pretrained InceptionV3 architecture, in which hyperparameter tuning and the image preprocessing were performed. The significance of the classification is that, results allow us to understand the impact of each class in the binary and multi class classification tasks. That is from the generated results we can analyse the performance and contribution of each class in the final classification task. This helps to get a better understanding over each class.

4. Results and discussion
The table 2 demonstrates the performance of Messidor data set on pretrained models- VGG19, InceptionV3, MobileNet, ResNet and NASNet. All the models achieved nearly similar accuracies ranging from 48.75% to 51.25%, but Depending on each model, the time taken to complete one epoch varies considerably, which was less for VGG19 and high for NASNet, due to the complexity of the model.

| Model      | Time taken to complete one epoch | Validation accuracy (%) |
|------------|----------------------------------|-------------------------|
| VGG19      | 13 sec                           | 48.83                   |
| InceptionV3| 63 sec                           | 48.75                   |
| MobileNet  | 111 sec                          | 51.25                   |
| ResNet     | 32 sec                           | 50.00                   |
| NASNet     | 121 sec                          | 50.62                   |

Binary classification of Messidor data set was carried out on three pretrained models (VGG19, ResNet, and InceptionV3) using three different optimizers (Adadelta, Adam and SGD). The results are tabulated in table 3. It was observed that by using Adam and SGD optimizers, validation and test accuracies were same and the model was classifying the result to only one class. Adadelta as optimizer could perform slightly better compared to other optimizers in all the models.

InceptionV3 was chosen as a pre-trained model for further experiments, since it follows the architecture of the analysis [3] and performs relatively better. Preprocessing was done using CLAHE algorithm and powerlaw transformation. The output of sample fundus image, the image after CLAHE preprocessing, and the image after powerlaw transformation on CLAHE is shown in figure 2, figure 3, and figure 4 respectively. The results of CLAHE pre-processing is laid out in table 4. Adam optimizer with batch size 32 gives comparatively better validation and testing accuracies.

Since Adam was giving comparatively better results for CLAHE algorithm, powerlaw transformation has been performed on CLAHE images. The results are depicted in table 5. Relatively better results were produced by using Adam optimizer with batch size of 32.
Table 3. Messidor data set on various pretrained architectures.

| Architecture | Optimizer | Batch size | Validation accuracy (%) | Test accuracy (%) |
|--------------|-----------|------------|-------------------------|------------------|
| VGG19        | Adadelta  | 16         | 50                      | 50               |
|              |           | 32         | 49                      | 45               |
|              | Adam      | 16         | 55                      | 55               |
|              |           | 32         | 55                      | 55               |
|              | SGD       | 16         | 46                      | 57               |
|              |           | 32         | 54                      | 55               |
| ResNet       | Adadelta  | 16         | 50                      | 55               |
|              |           | 32         | 55                      | 53               |
|              | Adam      | 16         | 55                      | 55               |
|              |           | 32         | 55                      | 55               |
|              | SGD       | 16         | 55                      | 55               |
|              |           | 32         | 53                      | 55               |
| InceptionV3  | Adadelta  | 16         | 56                      | 51               |
|              |           | 32         | 46                      | 48               |
|              | Adam      | 16         | 55                      | 55               |
|              |           | 32         | 55                      | 55               |
|              | SGD       | 16         | 44                      | 45               |
|              |           | 32         | 55                      | 55               |

Figure 2. Input image.

Figure 3. CLAHE Image.

Figure 4. Image after Powerlaw transformation
Hyperparameter tuning was performed by choosing different learning rates, batch sizes, momentum and activation functions. The learning rates chosen was 0.1, 0.01, 0.001 with Adam optimizer, which gave better results in previous experiments with different optimizers. The highest training and testing accuracy were produced by the learning rate 0.001. The results are present in Table 6.

### Table 6. Different learning rate on InceptionV3.

| Learning rate | Validation accuracy (%) | Test accuracy (%) |
|---------------|-------------------------|-------------------|
| lr=0.1        | 55                      | 45                |
| lr=0.01       | 55                      | 55                |
| lr=0.001      | 58                      | 50                |

Different batch sizes of 8, 16 and 32 were selected with learning rate 0.001 and Adam as optimizer. Comparatively better accuracies were obtained for batch size of 32. Other accuracy values are shown in Table 7.

### Table 7. Different batch sizes on InceptionV3.

| Batch size | Validation accuracy (%) | Test accuracy (%) |
|------------|-------------------------|-------------------|
| 8          | 55                      | 55                |
| 16         | 58                      | 50                |
| 32         | 58                      | 54                |

The table 8 presents the result using different momentum with Adam as optimizer, batch size of 32, and learning rate of 0.001. The values 0.9, 0.09 and 0.009 were chosen as different momentum. Better accuracies were obtained for momentum value 0.09.

### Table 8. Different momentums on InceptionV3.

| Momentum | Validation accuracy (%) | Test accuracy (%) |
|----------|-------------------------|-------------------|
| 0.9      | 47                      | 53                |
| 0.09     | 54                      | 53                |
| 0.009    | 46                      | 46                |

Finally, different activation functions were applied on InceptionV3 architecture with Adam optimizer, batch size of 32, learning rate of 0.001, and momentum of 0.09. The results are depicted in table 9.
Table 9. Different activation functions on Messidor dataset.

| Activation function | Validation accuracy (%) |
|---------------------|-------------------------|
| Sigmoid             | 55                      |
| Relu                | 55                      |
| Elu                 | 55                      |
| tanh                | 47                      |
| linear              | 47                      |
| selu                | 54                      |
| exponential         | 45                      |
| swish               | 48                      |

The novelty of the work is that none of the approaches in the literature have made any attempt to examine the inter class similarity of the DR images, which has crucial role in the performance of DL algorithms. The results of the classification between each class gives us the scope of identifying contribution of each class in the final learning process. Therefore, in order to identify impact of each class all the classes were tested with each other as delineated in table 10. The experiment also was carried out using InceptionV3 architecture and other hyperparameters. The results show that the model was unable to differentiate samples as it was only classifying input images to a single class. From the results, it was observed that class zero and class two have predominance over the other the classes. So, while performing the binary class and multi class classification results from other classes have to be evaluated keenly since they have subtle features.

Table 10. Inter class classification result

| Classes       | Validation accuracy (%) | Test accuracy (%) | Remarks                  |
|---------------|-------------------------|-------------------|--------------------------|
| 0 and 1       | 78                      | 78                | Classifying all classes to zero |
| 0 and 2       | 69                      | 69                | Classifying all classes to zero |
| 0 and 3       | 32                      | 32                | Classifying all classes to three |
| 1 and 2       | 61                      | 61                | Classifying all classes to two |
| 1 and 3       | 62                      | 62                | Classifying all classes to three |
| 2 and 3       | 49                      | 49                | Classifying all classes to two |

5. Conclusion and Future scope

Computer aided methods for DR screening has become prevalent in the recent years. Many image processing and Machine Learning (ML) algorithms have been applied to standard datasets for DR detection and grading. The study analyses performance of CNN architectures on binary class, multi class and inter class classification of DR. The deeper level hyper parameter tuning could improve the accuracy of the model but, only to some extent. We could also observe that the performance of the model decreases with raise in the number of classes. In inter class classification, the features that distinguish severe vs normal class show very poor results indicating that, level of subtleness is very high. The performance of class 0 and class 2 over class 3 shows the least accuracy values. The results indicate that binary class and multi class classification of DR has to be made after resolving interclass performances. Other common draw backs like generalizability of the model, understanding of how the neural networks make predicted output etc., has to be taken care before widely deploying the model.

References
[1] https://idf.org/our-network/regions-members/south-east-asia/members/94-india.html
[2] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. and Kim, R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.
[3] Kanagasingam, Y., Xiao, D., Vignarajan, J., Preetham, A., Tay-Kearney, M.L. and Mehrotra, A., 2018. Evaluation of artificial intelligence–based grading of diabetic retinopathy in primary care. *JAMA network open*, 1(5), pp.e182665-e182665.

[4] Ishtiaq, U., Kareem, S.A., Abdullah, E.R.M.F., Mujtaba, G., Jahangir, R. and Ghafoor, H.Y., 2020. Diabetic retinopathy detection through artificial intelligent techniques: A review and open issues. *Multimedia Tools and Applications*, 79(21), pp.15209-15252.

[5] Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B. and Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE transactions on medical imaging*, 35(5), pp.1299-1312.

[6] Badar, M., Haris, M. and Fatima, A., 2020. Application of deep learning for retinal image analysis: A review. *Computer Science Review*, 35, p.100203.

[7] Sahlsten, J., Jaskari, J., Kivinen, J., Turunen, L., Jaanio, E., Hietala, K. and Kaski, K., 2019. Deep learning fundus image analysis for diabetic retinopathy and macular edema grading. *Scientific reports*, 9(1), pp.1-11.

[8] Ram, P. and Swapna, T.R., 2018. A Study on Various Quantification Algorithms for Diabetic Retinopathy and Diabetic Maculopathy Grading. *Computational Vision and Bio Inspired Computing* (pp. 412-421).

[9] Kiruthika, M., Swapna, T.R., Santhosh, K.C. and Peeyush, K.P., 2019, April. Artery and Vein classification for hypertensive retinopathy. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), (pp. 244-248).

[10] Lam, C., Yi, D., Guo, M. and Lindsey, T., 2018. Automated detection of diabetic retinopathy using deep learning. *AMIA summits on translational science proceedings, 2018*, p.147.

[11] Setiawan, W., Utoyo, M.I. and Rulaningtyas, R., 2019. Classification of neovascularization using convolutional neural network model. *TELKOMNIKA*, 17(1), pp.463-472.

[12] Sujee, R. and Padmavathi, S., 2017, January. Image enhancement through pyramid histogram matching. 2017 International Conference on Computer Communication and Informatics (ICCCI), (pp. 1-5).

[13] Achan, E.R. and Swapna, T.R., 2018, May. Hidden Markov Random Field and Gaussian Mixture Model Based Hidden Markov Random Field for Contour Labelling of Exudates in Diabetic Retinopathy—A Comparative Study. *International Conference on ISMAC in Computational Vision and Bio-Engineering* (pp. 1307-1317).

[14] de La Torre, J., Valls, A. and Puig, D., 2020. A deep learning interpretable classifier for diabetic retinopathy disease grading. *Neurocomputing*, 396, pp.465-476.

[15] Hagos, M.T. and Kant, S., 2019. Transfer learning based detection of diabetic retinopathy from small dataset. *arXiv preprint arXiv:1905.07203*.

[16] Ahmad, J., Sajjad, M., Mehmood, I. and Baik, S.W., 2017. SiNC: Saliency-injected neural codes for representation and efficient retrieval of medical radiographs. *PloS one*, 12(8), p.e0181707.

[17] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826.