Assessment of transfer learning models for grading of diabetic retinopathy

Sowmiya M¹, Banu Rekha B² and Malar E³

Abstract
Diabetic retinopathy is a potentially mortal diabetic complication. The severity level of DR must be identified earlier to reduce the medical complications. Effective automated ways for identifying DR and classifying its severity stage are necessary to reduce the burden on ophthalmologists. Transfer learning methods are utilized to automatically grade the severity of diabetic retinopathy in this study. The stages of DR are diagnosed using pretrained VGG16, Inception v3, and ResNet50 models on pre-processed retinal images of DDR dataset. Out of three implemented models, Inception v3 achieved higher validation accuracy of 76.47% and testing accuracy of 67% compared to VGG16 and ResNet50 models. This research contributes to the analysis of deep learning architectures for the creation of automated diabetic retinopathy stage diagnosis and grading.

Keywords: Transfer learning, retinal image, diabetic retinopathy, VGG16, Inception v3, ResNet50.

Introduction
Diabetic retinopathy (DR), which primarily affects working people, is the leading cause of vision impairment and blindness. Diabetes causes diabetic retinopathy, which is caused by elevated blood sugar levels. This medical disorder causes vascular damage and aberrant blood vessels (Duh et al., 2017). The International Diabetes Federation predicts that there will be 643 million diabetic people globally in 2030 and 783 million by 2045 (Ogurtsova et al., 2017). Nonproliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR) are two stages of DR. NPDR manifests itself in the early phases, which are further divided into Mild, Moderate, and Severe. Microaneurysms (MA), dot and flame shaped haemorrhages, exudates, and cotton wool patches are all common NPDR abnormalities (Duh et al., 2017). PDR is a more highly developed form of DR in which aberrant vessels begin to form and lead to neovascularization. In addition to lesions, vascular changes include vessel calibres, tortuosity, and branching angles (Habib et al., 2014). The mild stage might progress to advanced PDR if an early diagnosis is not made. Early screening and treatment are required to limit retinal complications and avoid vision loss. Because manual grading needs skill and time, Computer Assisted Diagnosis (CAD) techniques have been demonstrated to be helpful in diagnosing the DR at an early phase. Table 1 shows the abnormal lesions appear in various stages of DR.

Several deep learning-based techniques for diagnosing DR using fundus images have been proposed. CNN models were created specifically to rate the severity degree (Alyoubi et al., 2020). However, the DL model necessitates a vast amount of data. Because the current database has a restricted number of photographs. Transfer learning strategies have been highlighted in studies as a way to overcome restricted resources (Oltu et al., 2021). The pre-trained CNN models were utilized specifically for assessing the severity level of DR. To diagnose the severity of DR, pretrained architectures Alexnet, VGG16, VGG19, Inceptionv3, ResNet were used (Oltu et al., 2021) . The Kaggle EyePACS, Messidor, DIARETDB, and e-optha databases were used in the majority of the studies (Alyoubi et al., 2020). The EyePACS dataset comprises...
around 88,000 images, which is enough data for DL models. However, it has the drawback of including ungraded and low-quality photos (Alyoubi et al., 2020). In this study, a second large diabetic retinopathy dataset DDR is used, and the performance of transfer learning models is examined. The sample images for stages of DR taken from the DDR dataset are shown in Figure 1.

Employed DDR datasets, hence this work focuses on constructing transfer learning models for DDR datasets.

Pre-processing, selecting a pretrained architecture, and fine-tuning the architecture are the primary processes necessary for detecting DR using a transfer learning approach. Riaz et al. classified DR into five groups using the CNN architecture. The model was created with the help of the kaggle and messidor datasets. Using the EyePACS and Messidor-2 datasets, Gulshan et al. developed an Inception-v3 model to predict diabetic retinopathy. The model had a sensitivity of 90.3% for the EyePACS dataset and 87% for the Messidor-2 dataset (Riaz et al., 2020). Data augmentation was used in conjunction with CNN to diagnose five categories of DR severity levels, with a sensitivity of 95% (Chlap et al., 2021). On the Kaggle dataset, Kanungo et al. built the Inceptionv3 architecture for automated detection of various of diabetic retinopathy. For a batch size of 128, they were able to attain an accuracy of 88% (Kanungo et al., 2017). Masood et al. used CNN transfer learning with Inceptionv3, which had been pretrained on ImageNet. They were able to attain 48.2% accuracy on the EyePACS five class dataset Masood et al., 2017). For the diagnosis of diabetic retinopathy, Mansour et al. created AlexNet architecture with various optimization approaches (Mansour, 2018). On a dataset containing 30,000 photos, Gosh et al. used a CNN-based model to achieve an accuracy of 95 and 85% for two and five category diabetic retinopathy disorders (Ghosh et al., 2017). Chandrakumar et al. devised the stagewise classification model, which includes dropout to reduce overfitting. The model was developed using the Kaggle dataset and tested using the DRIVE and STARE datasets (Chandrakumar & Kathirvel, 2016). Using the EyePACS dataset, Wang et al. created AlexNet, VGG16, and Inception-v3 models. The accuracy rates were 37.43, 50.3, and 63.23%, respectively (Wang et al., 2018). Wan et al. applied pretrained VGG, AlexNet, ResNet, and GoogLeNet models to a dataset of 35,126 images from the

<table>
<thead>
<tr>
<th>Stages of DR</th>
<th>Retinal anomalies</th>
</tr>
</thead>
<tbody>
<tr>
<td>No DR</td>
<td>No anomalous signs</td>
</tr>
<tr>
<td>Mild NPDR</td>
<td>Few Microaneurysms</td>
</tr>
<tr>
<td>Moderate NPDR</td>
<td>Microaneurysms, hemorrhages, Exudates</td>
</tr>
<tr>
<td>Severe NPDR</td>
<td>venous beading, intraretinal microvascular abnormality</td>
</tr>
<tr>
<td>Proliferative DR</td>
<td>neovascularization, vitreous and preretinal hemorrhages</td>
</tr>
</tbody>
</table>

According to the literature, transfer learning is extensively utilized to diagnose the DR severity level for EyePACS Kaggle and MESSIDOR datasets. Only a few research have The contributions of this work are,

- This work develops the transfer learning pretrained models VGG16, Inception v3, ResNet50 for DDR dataset. From the literature, it is inferred that few studies have used this dataset.
- For grade five stages of DR, the DDR database will be used for training and validation. Performance criteria such as accuracy, sensitivity, precision, and F1 score will be examined for each architecture.
- This study evaluates the performance of three models for DDR dataset.

The paper is organized as follows; Section 2 presents the description of dataset and section 3 presents the methods and transfer learning models. Experimental results and discussions are presented in section 4 followed by a conclusion.

**Dataset**

The DDR dataset contains 13,637 fundus pictures showing evidence of DR with a resolution of 2124 x 2056 and a FOV.
Transfer learning models for diabetic retinopathy

Images are collected between 2016 to 2018 in China from 147 hospitals. The level of DR severity includes No DR, mild NPDR, moderate NPDR, severe NPDR, and PDR. Images are graded by multiple graders. The dataset contains training, validation, and testing sets separately (Li et al., 2019).

Methodology
The steps in this work are depicted in Figure 2. Pre-processing, training, and classification with performance metrics evaluation are the three stages of the transfer learning method given in this paper. The images were first resized and normalized as part of the pre-processing procedure as shown in Figure 3. The data augmentation was used to reduce overfitting and improve the model's generalization ability (Tariq et al., 2021). Three CNN models were employed in the training stage: VGG16, Inception v3, and ResNet50, all of which were pre-trained using the ImageNet database. The last layers of each model were fine-tuned, and a dense layer was added for grading.

Transfer Learning Models
In this study, pre-trained CNN architectures, VGG16, Inception v3, and ResNet50, were trained and validated on the DDR database. For DR grading, fine-tune the architecture’s final fully connected layers. The following is a model training procedure:
- Resize, normalize and augment the data in the training set as part of the pre-processing.
- Load the models that have been pre-trained.
- Configure the trained model’s parameters, such as batch size, epochs, optimizer, and metrics.
- Save the model and run it.
- Evaluate the model’s loss and other performance measures.

VGG16
VGG16 is a CNN model developed for the 2014 ImageNet Challenge by the Visual Geometry Group (Simonyan & Zisserman, 2014). There are 13 convolution layers, 5 pooling layers, and 3 fully connected layers in the VGG16 network. The image is resized to a size of 224x224. The model is pre-trained using the ImageNet database and weights are

<table>
<thead>
<tr>
<th>Class</th>
<th>DR stage</th>
<th>Total</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No DR</td>
<td>6266</td>
<td>3133</td>
<td>1253</td>
<td>1880</td>
</tr>
<tr>
<td>1</td>
<td>Mild</td>
<td>630</td>
<td>315</td>
<td>126</td>
<td>189</td>
</tr>
<tr>
<td>2</td>
<td>Moderate</td>
<td>4477</td>
<td>2238</td>
<td>895</td>
<td>1344</td>
</tr>
<tr>
<td>3</td>
<td>Severe</td>
<td>236</td>
<td>118</td>
<td>47</td>
<td>71</td>
</tr>
<tr>
<td>4</td>
<td>Proliferative</td>
<td>913</td>
<td>456</td>
<td>182</td>
<td>275</td>
</tr>
<tr>
<td>5</td>
<td>Ungradable</td>
<td>1151</td>
<td>575</td>
<td>230</td>
<td>346</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13,673</td>
<td>6835</td>
<td>2733</td>
<td>4105</td>
</tr>
</tbody>
</table>

Figure 2: Framework for grading the DR severity level

Table 2: Shows the DR grading distribution

Pre-processing (Image Resize, Data augmentation, normalization)
Transfer Learning models VGG16, Inceptionv3, ResNet (Pretrained with ImageNet database)
Fine-tuning and modifying the fully connected layers
DR grading and evaluate the performance metrics

VGG16
VGG16 is a CNN model developed for the 2014 ImageNet Challenge by the Visual Geometry Group (Simonyan & Zisserman, 2014). There are 13 convolution layers, 5 pooling layers, and 3 fully connected layers in the VGG16 network. The image is resized to a size of 224x224. The model is pre-trained using the ImageNet database and weights are
Inception v3
The Inception v3 is an enhanced version of the Inception v1, which was first released as GoogLeNet in 2014. Convolutional layers, pooling layers with an inception block, and fully connected layers are among the 48 layers in Inception v3. The Inception modules combine the characteristics of preceding levels to speed up calculation. Batch normalization and dropout of 0.5 were employed to the design to boost performance (Szegedy et al., 2015). The model allows images with a resolution of 299x299. In this work, pretrained layers with fully connected layer and SoftMax layer are employed for classification.

ResNet50
ResNet50, also known as Residual Network, is a 50-layer residual network with skip connections that reduces the gradient problem (He et al., 2015). The architecture begins with a 7x7 convolution layer, which is followed by a maxpooling layer. Four stages with three residual blocks are layered, each with three convolution layers of kernel size 1x1, 3x3, 1x1. Finally, DR is graded using a 0.5 dropout, fully connected layer, and SoftMax layer. Figures 5 and 6 illustrate the Inception v3 and ResNet50 architecture.

RESULTS AND DISCUSSIONS

Performance Metrics
Accuracy, sensitivity, precision, and F1 score are utilized as performance metrics to quantitatively evaluate the model presented in Table 4.

where, TP – True Positive; TN – True Negative; FP – False Positive; FN – False Negative.

The percentage of correctly categorized samples is called accuracy. Precision is the fraction of accurately classified samples among the classified positive samples. The fraction of samples accurately categorized from all positive samples is measured by recall.

Results of Transfer Learning Models
This section contains the setup and outcomes for all of the pretrained models, as well as the loss gained during training and each model’s class-wise performance. VGG16, Inception v3, and ResNet50 are the transfer learning architectures employed. For DR grading, each model employed pretrained weights with modified last layers. Six
Transfer learning models for diabetic retinopathy

class targets are added to the SoftMax layer. The pretrained models are fine-tuned using a 6835-image training dataset, validated using 2733 photos, and tested with 4105 images. The model’s hyperparameters are optimized to maximize the results displayed in Table 3. The Adam optimizer with sparse categorical cross entropy loss is used for all of the architectures. Table 5 presents the accuracy of all the models. Training, validation, and testing accuracy have all been presented.

In comparison to the other two models, the inception v3 model gave better outcomes, as seen in Table 5. Furthermore, the training and validation accuracy have not been varied with a considerable variation, indicating that the overfitting problem has been reduced. Validation accuracy of 76.47% and testing accuracy of 67% were attained by Inception v3.

When trained on the training and validation datasets, the accuracy and loss performance of the three models is shown in Figure 7. Because the models were resistant to overfitting, they performed well during training and validation. Over each epoch, the losses in training and validation have decreased.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training (%)</th>
<th>Validation (%)</th>
<th>Testing (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16</td>
<td>55.92</td>
<td>65.85</td>
<td>57</td>
</tr>
<tr>
<td>Inceptionv3</td>
<td>68.01</td>
<td>76.47</td>
<td>67</td>
</tr>
<tr>
<td>ResNet50</td>
<td>42</td>
<td>51.9</td>
<td>53.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No DR (0)</td>
<td>0.56</td>
<td>0.97</td>
<td>0.71</td>
</tr>
<tr>
<td>Mild DR (1)</td>
<td>0.11</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>Moderate DR (2)</td>
<td>0.36</td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>Severe DR (3)</td>
<td>0.1</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Proliferative DR (4)</td>
<td>0.52</td>
<td>0.20</td>
<td>0.29</td>
</tr>
<tr>
<td>Ungradable DR (5)</td>
<td>0.80</td>
<td>0.38</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Figure 5: Inception v3 architecture for DR Grading

Figure 6: ResNet50 architecture for DR Grading
The metrics collected for each class for the DDR database are presented in Tables 6-8. Because of the enormous sample size, models perform better in the no DR and ungradable classes. It demonstrates that the models were successful in categorizing these two classes. However, all the models fail to account for mild DR; specifically, VGG16 have not identified the severe DR class. The achieved results shows that the models were identified each class with better performance. Of all three models, Inception v3 produced the highest validation accuracy of 76.47% and testing accuracy of 67%, respectively. Inception blocks produce better results for DR classification than VGG16 and ResNet50 architectures.

**Conclusion**

Transfer learning-based diagnosis models for grading diabetic retinopathy are presented in this research. Three pre-trained CNN models were used, and the final layers for multi-class grading were fine-tuned. Because the bulk of the class belongs to the no DR, there may be a lot of bias. Dropout and data augmentation were used to reduce bias and overfitting. The performance of VGG16, Inception v3, ResNet50 models have been analyzed for DDR dataset. From the results, the Inceptionv3 model achieved better performance with an accuracy of validation accuracy of 76.47% and testing accuracy of 67%. VGG16 produced a validation accuracy of 65.87% and testing accuracy of 57%. ResNet50 produced a validation accuracy of 51.9% and a testing accuracy of 53.6%. The research findings could be used to larger datasets in the future to increase the performance. Additionally, models will be trained from scratch and analyze its performance for clinical implementation.

**References**


Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., & Kang, H. (2019). Diagnostic assessment of deep learning algorithms for diabetic...


