Is Diabetic Retinopathy Grading Biased by Imbalanced Datasets?

Fernando C. Monteiro[0000−0002−1421−8006] and José Rufino[0000−0002−1344−8264]

Research Centre in Digitalization and Intelligent Robotics (CeDRI),
Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC)
Instituto Politécnico de Bragança,
Campus de Santa Apolónia, 5300-253 Bragança, Portugal
monteiro,rufino@ipb.pt

Abstract. Diabetic retinopathy (DR) is one of the most severe complications of diabetes and the leading cause of vision loss and even blindness. Retinal screening contributes to early detection and treatment of diabetic retinopathy. This eye disease has five stages, namely normal, mild, moderate, severe and proliferative diabetic retinopathy. Usually, highly trained ophthalmologists are capable of manually identifying the presence or absence of retinopathy in retinal images. Several automated deep learning (DL) based approaches have been proposed and they have been proven to be a powerful tool for DR detection and classification. However, these approaches are usually biased by the cardinality of each grade set, as the overall accuracy benefits the largest sets in detriment of smaller ones. In this paper, we applied several state-of-the-art DL approaches, using a 5-fold cross-validation technique. The experiments were conducted on a balanced DDR dataset containing 31330 retina fundus images by completing the small grade sets with samples from other well known datasets. This balanced dataset increases robustness of training and testing tasks as they used samples from several origins and obtained with different equipment. The results confirm the bias introduced by using imbalanced datasets in automatic diabetic retinopathy grading.

Keywords: Diabetic retinopathy grading · Deep learning network · Retinal fundus images · Diabetic retinopathy dataset · Imbalanced dataset.

1 Introduction

Diabetic Retinopathy is a serious and common health condition, caused by diabetes mellitus, that affects the human retina, causing damage to blood vessels which become leaky or blocked, generating microaneurysms, soft and hard exhudates and haemorrhages [29] (see Fig. 1). Blood vessels grow and swelling in the central part of the retina can lead to vision loss or even blindness.

Although there is no knowledge of the total number of people affected with moderate and severe forms of glaucoma or diabetic retinopathy, it is estimated that several million cases of eye diseases could have been avoided if there had been a timely diagnosis [34]. It is known that screening, early detection and

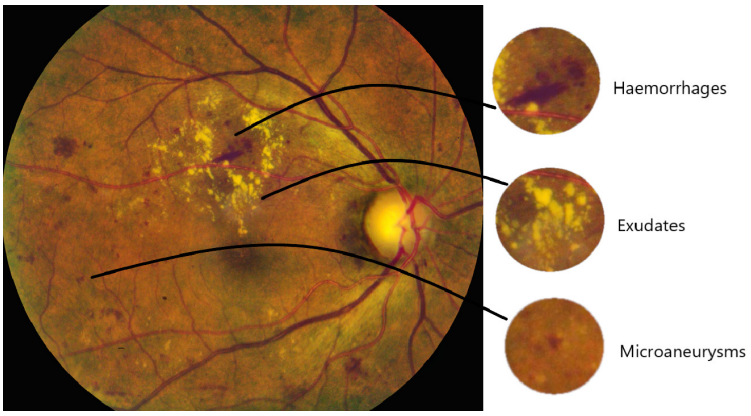


Fig. 1. DR lesions in a retina fundus image.

prompt treatment of DR allow prevention of visual impairment. With the rising incidence and prevalence of DR, public health systems in both developing and developed countries should implement and maintain a DR screening program for people with diabetes [30].

Accordingly with the International Clinical Diabetic Retinopathy Disease Severity Scale [33] (ICDRDSS), there are five severity levels of DR depending on the presence of retina lesions, namely: *normal* (grade 0) - no abnormalities; *mild DR* (grade 1) - tiny bulges develop in the blood vessels, which may bleed slightly but do not usually affect vision; *moderate DR* (grade 2) - swelling and distortion of blood vessels; *severe DR* (grade 3) - widespread changes affect the blood vessels, blocking them, including more significant bleeding into the eye and prominent microvascular abnormalities; *proliferative DR* (grade 4) - neovascularization or vitreous/preretinal hemorrhage which can result in loss of vision. This grading is important to determine the exact stage of DR to begin a suitable treatment process and to prevent the retina deterioration and the possible blindness.

A worldwide problem associated with DR is the reduced number of specialists to follow the patients: e. g., in China there is a ratio of one ophthalmologist for 3000 diabetes patients [8]. To overcome this problem, several computer aided diagnosis techniques, mainly based on deep learning schemes, have been proposed to automatically detect DR and its different stages from retina fundus images.

Deep learning models are designed to automatically learn feature hierarchies through back-propagation by using multiple layer blocks from low to high level patterns [4]. The automatic extraction of the most discriminant features from a set of training images, suppressing the need for preliminary feature extraction, became the main strength of deep learning approaches.

There still remain some concerns associated with the automated DR classification approaches proposed in the literature that are preventing their use in clinical applications. Although such methods have achieved high performance in a specific imbalanced dataset, more studies are need to known how they may

perform on other datasets. Here, the major issue at stake is the reduced model reliability. An important underlying factor is the fact that the majority of published approaches used imbalanced datasets in their studies. Using a dataset where more than 70% of the samples are healthy retina images and only 2% for one of DR grades will introduce bias in the results. Alyoubi et al. in their review paper [2] discuss the existing DR grading techniques and have shown that even when doing multi-level classification, the available DL based approaches provide only a DR grading result.

Deep learning networks (DLN) automatically extract the most discriminative features from training images. Nevertheless, which features are extracted to make a classification decision remains somehow unknown. Since experts usually evaluate the DR grade based on the presence of lesions in the retina, we believe that while automated techniques do not demonstrate that their results are based on the lesions, such techniques will not be accepted as a reliable substitute of human ophthalmologists. To sum up, more and thorough investigations about the usage of deep learning methods work for DR grading are needed.

In this paper we introduce a deep learning-based method for the problem of classifying DR in fundus imagery. To avoid over-fitting produced by imbalanced datasets in one DLN, our approach applies several of the best known DLN models that were trained and tested in a balanced dataset. This balanced dataset increases the robustness of the model, allowing it to produce more reliable predictive results than those obtained by using imbalanced datasets.

The remainder of this paper is organized as follows: in Section 2, previous related works are presented and reviewed; in Section 3, we describe the used materials and the proposed method; Section 4 presents the results and some discussion on the findings; finally, in Section 5, some conclusions are drawn.

2 Related Works

Automatic systems for DR recognition based on image features have been proposed for a long time. In general terms, those approaches extract some feature characteristics to identify retina lesions [5]. The majority of these classification techniques are binary classifiers that classify data only into two classes: normal and DR affected [13]. The problem with these methods is that they do not allow to identify the disease severity even when it is known that a patient has DR.

Recent increasing of computational resources and capabilities have created the opportunity to develop deep learning models for DR detection and grading. This section reviews the studies in which the DR datasets (EyePACS, APTOS or DDR) were classified into five classes accordingly with the ICDRDSS scale. The majority of published DR grading studies used imbalanced datasets.

Pratt et al. [21] presented one of the first studies employing a DLN for classifying DR samples in five classes. They obtained an accuracy of 0.75 but could not classify the mild class accurately. An important limitation is that they trained the network with limited number of images from a skewed EyePACS dataset, which could prevent the DLN from learning more features.

Wan et al. [32] studied the performance of four available pretrained networks to detect the five stages in the EyePACS dataset. During the preprocessing stage, the images were filtered with a NonLocal Means Denoising function. Transfer learning was done by fine tuning the last fully connected layer. They reported the best accuracy of 0.957 for the VGG16 architecture. However, they have used the biased imbalanced EyePACS dataset.

Bodapati et al. [6] presented an approach where features extracted from multiple pre-trained DLN are blended using a multi-modal fusion module. These final weights are used to train a DLN employed in DR grading prediction. Experiments were conducted in the APTOS dataset using 80% for training and 20% for validation. The authors did not used a test dataset. They achieved an accuracy of 0.81 over the validation dataset. Unfortunately, they have used the imbalanced version of the APTOS dataset.

Li et al. [17] classified DR levels in their dataset (DDR) by fine-tuning five available pretrained networks. They identified a new DR grade for images without quality for being classified. The SE-BN-Inception network [15] obtained the best accuracy of 0.828, including grade 5 samples, and 0.751, removing grade 5 samples. The work highlights the difficulty in classifying mild DR, and the frequency with which severe DR is misclassified as moderate. Still, they have used an imbalanced version of the DDR dataset, that produces biased results.

Majumder and Kehtarnavaz [19] proposed a multitasking deep learning model for DR grading in five levels. They used a Multitasking Squeeze Excitation Densely Connected Network that consists of a SEDenseNet classification model, a SEDenseNet regression model, and a MLP classifier. They achieved an accuracy of 0.82 for the EyePACS dataset and 0.85 for the APTOS dataset. Unfortunately, they have used the imbalanced version of the datasets.

Alyoubi et al. [1] proposed a deep learning model (CNN512) to classify retina fundus images into one of the five DR stages. To overcome the different levels of light in the samples, the retina images were normalised. For the normalization, they subtract the mean and then divide the variance of the images. This achieved an accuracy of 0.886 and 0.841 on the DDR and the APTOS public datasets, respectively. Again, the imbalanced version of the datasets were used.

Some studies [22,24] identified the problem associated with imbalanced datasets and achieved balance among grade sets by replicating a small number of images to a large number by data augmentation. In spite of that, the variability of features trained and tested will be very low, introducing bias in the results.

Qummar et al. [22] ensembled five DLN models to encode the best features and improve the classification for different stages of DR. They used stacking to combine the results of all different models and generate a unified output. They obtained an overall accuracy of 0.648. To overcome the bias created by the imbalanced EyePACS dataset they balanced it via data augmentation. However, this process used only the original images in each grade set. For example, in the grade 4 set they obtained 4119 images from just 113 original images. The variability of features in such augmented images is, of course, very small.

In a recent work, Rocha et al. [24] applied the VGG16 network to classify the diabetic retinopathy in five and six levels in the DDR, EyePACS and IDRiD databases. They balanced the three datasets by using data augmentation over the original images as in [22], thus with the same variability issue. The number of samples per class was 800 (DDR), 3000 (EyePACS) and 140 (IDRiD). The accuracy for five levels in each dataset was 0.854, 0.779 and 0.898, for the DDR, EyePACS and IDRiD datasets, respectively.

For comprehensive details of research in the field of DR detection readers are encouraged to see the excellent reviews of Alyoubi et al. [2] and Tsiknakis et al. [31]. In these works the authors discuss the advantages and disadvantages of current DR classification techniques and analyse public DR datasets available.

3 Proposed Method

In this section we introduce the details of our framework. First, we introduce the DR datasets, then the preprocessing methods and, finally, the DL approach to classify the samples in the five DR stages.

3.1 Diabetic retinopathy dataset

The automation of DR recognition and grading depends on large image datasets with many samples categorized by ophthalmologists. The results depend on the number of samples and their acquisition conditions. A small number of images may result in poor learning models, that are not sufficient to adequately train the deep learning architecture.

Another factor to consider in DL approaches is the size and quality of datasets used for training and testing datasets. Over-fitting is a major issue in DLN, as imbalanced datasets make the network to over-fit to the most prominent class in the dataset [31].

Most of the publicly available datasets contain less than 2000 images, like the Indian Diabetic Retinopathy Image dataset (IDRiD) [20], with 516 images, or Messidor 2 [9], with 1748 images. The Asia Pacific Tele-Ophthalmology Society (*Kaggle* APTOS) [3] dataset contains 5590 images (3662 samples for training and 1928 samples for testing) collected by the Aravind Eye Hospital in India; however, only the ground truths of the training images are publicly available.

Kaggle EyePACS [12] is the largest DR dataset with 88702 images (35126 samples for training and 53576 for testing) classified into five DR levels. This dataset was classified by only one ophthalmologist, which can potentially lead to annotation bias. It consists of a large number of images which were obtained under a variety of imaging conditions by different devices at several primary care sites throughout California [12]. Maybe due to such variability, there is a large number of images where the eye's features are indistinguishable due to the presence of noise, artifacts produced by lenses, chromatic distortions or low light. In fact, accordingly with the dataset curators, the presence of these extraneous features was an intended goal, in order to better simulate a real world scenario.

The DDR dataset [17] is the second largest dataset, consisting of 13673 images divided in 6835 for training, 2733 for validation and 4105 for testing. Retina fundus images were collected between 2016 and 2018 across 147 hospitals in China, and annotated by several ophthalmologists according to the ICDRDSS scale, using a majority voting approach.

DR screening produces some low quality images that are meaningless to analyze and grade due to the lack of information. In this way, a new class (Grade 5) was created to include these ungradable images, as showed in Fig. 2.

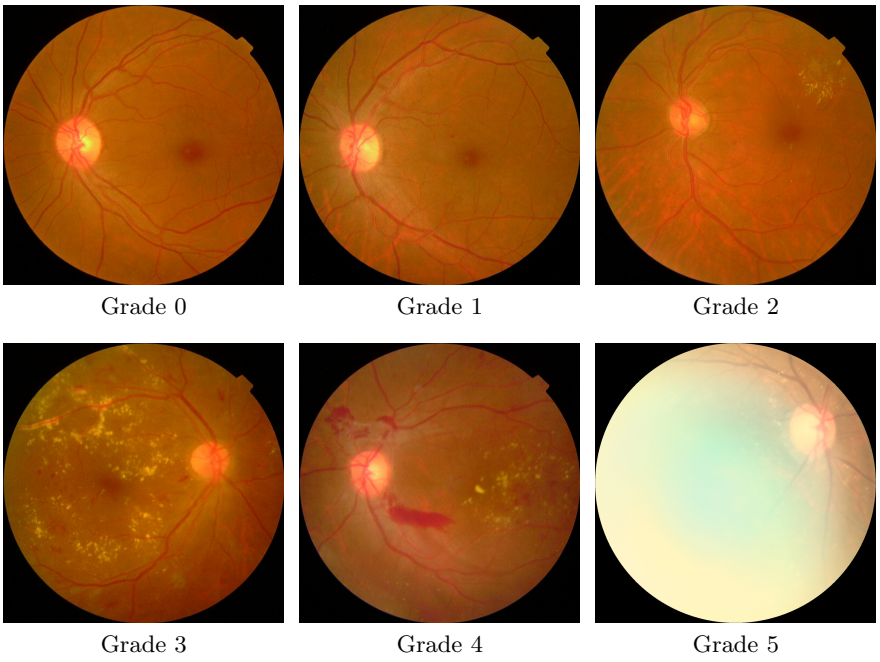


Fig. 2. The DR grades: normal retina (Grade 0), mild DR (Grade 1), moderate DR (Grade 2), severe DR (Grade 3), proliferative DR (Grade 4), ungradable (Grade 5).

Table 1 shows the distribution of training, validation and test images per class, for DR grading in the two largest publicly available datasets (EyePACS and DDR). The training set of the EyePACS dataset is usually split in 80% for training and 20% for validation [21].

The class labels of the datasets are highly imbalanced, specially in the *Kaggle* EyePACS dataset, with more than 73% of the samples annotated as normal, whereas severe and proliferative DR samples only account for less than 3% each.

To reduce the classification bias, Li et al. [17] and Qummar et al. [22] have augmented existing samples from the 1, 2, 3, and 4 classes, to produce new samples. Although augmenting the dataset does not make it totally balanced, it does help to lessen the biases. Nevertheless, as this augmentation process, in

Table 1. Number of samples per class in the training, validation and test sets of largest publicly available DR datasets.

Dataset	DR class	Training	Validation	Test	Total	Percentage
EyePACS	Grade 0	25810	-	39533	65343	73.6%
	Grade 1	2443	-	3763	2329	7.0%
	Grade 2	5292	-	7860	13152	14.8%
	Grade 3	873	-	1214	2087	2.4%
	Grade 4	708	-	1206	1914	2.2%
	Total	35126	-	53576	88702	
DDR	Grade 0	3133	1253	1880	6266	45.8%
	Grade 1	315	126	189	630	4.6%
	Grade 2	2238	895	1344	4477	32.8%
	Grade 3	118	47	71	236	1.7%
	Grade 4	456	182	275	913	6.6%
	Grade 5	575	230	346	1151	8.5%
	Total	6835	2733	4105	13673	

some classes, is based only in a very small number of images, the variability of features learned by the DLN does not increase.

To overcome this problem, in our experiment we have balanced the DDR classes in all the sets, using random samples of the same class, from the EyePACS, APTOS and IDRiD datasets, thus obtaining the balanced DDR dataset (BDDR) with a total of 31330 samples. Usually, the works that have used the DDR dataset have removed the class 5 set. In order to do a fair comparative study, we also removed this class.

Table 2 shows the distribution of training, validation and test sets per class for the balanced DDR dataset. Only a small number of samples from grades 3 and 4 of the training set have been obtained by rotation, though even when joining all the five datasets we cannot obtain the number of samples we need.

Using several datasets, that were collected on different resolution, equipment, and geographies, allow to incorporate this variability into our model, which increases the generalization by reducing sensitivity to the image acquisition system. Using only one imbalanced dataset oversimplifies the identification process, making it impractical to be used for recognizing DR in images obtained worldwide in different conditions [31].

3.2 Image preprocessing

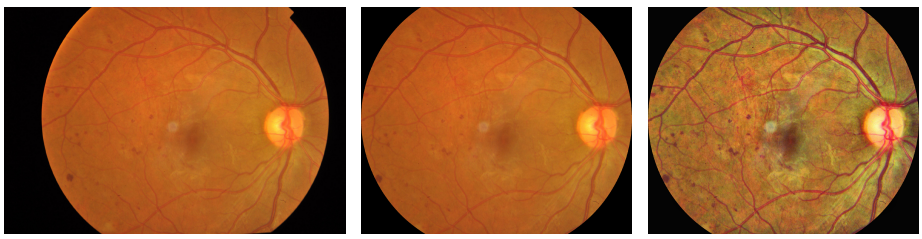
As the used DR datasets were obtained from different locations with different types of equipment, they may contain images that vary in terms of resolution and aspect ratio. The images could also contain uninformative black space areas that may induce some bias in the training and test processes.

Table 2. Number of samples in each grade set for training, validation and test sets of the balanced DDR dataset.

DR class	Training	Validation	Test	Total	Percentage
Grade 0	3133	1253	1880	6266	20.0%
Grade 1	3133	1253	1880	6266	20.0%
Grade 2	3133	1253	1880	6266	20.0%
Grade 3	3133	1253	1880	6266	20.0%
Grade 4	3133	1253	1880	6266	20.0%
Total	15665	6265	9400	31330	

Image preprocessing is thus a necessary step to enhance image features and to ensure the consistency of images. In order to reduce the black space areas, the images were cropped to a circle that limits the retina. The image resolution was also adapted to each deep learning architecture, e.g. 224x224 for ResNet and 299x299 for Inception architectures.

In some works only the green channel of images was extracted due to its high contrast in blood vessels and red lesions [18]. However, these approaches loose the colour information. Contrast enhancement is a common preprocessing approach used to highlight the image features like blood vessels or retina lesions. A technique often used is image normalization based on the min-pooling filtering [13,35]. In our experiments we adopted the Contrast Limited Adaptive Histogram Equalization (CLAHE), applying it to each colour channel. Figure 3 shows one original image and the outcome of the preprocessing techniques employed in it.



(a) (b) (c)

Fig. 3. Preprocessing steps: (a) original image, (b) cropped image, (c) CLAHE image.

3.3 Deep learning architectures

In our approach, ten state-of-the-art DLNs were used: VGG 16/19 [25], ResNet 50/101 [14], Inception-V3 [27], Inception-ResNet [26], Xception [7], DenseNet201 [16], DarkNet53 [23] and EfficientNetB0 [28].

Training a deep learning network is very demanding in terms of computational resources and data required. Unlike images of everyday objects, as those in the ImageNet dataset [11], large medical images datasets are very hard to obtain due to the necessary annotation time and legal issues involved [31].

Transfer learning is a commonly used technique that transfers the knowledge from a pretrained network on a large dataset and it is used to solve a recognition or classification task in which limited training data are available. All the ten deep learning architectures were fine-tuned by initialising their weights with ImageNet values and re-training all of their layers with the BDDR dataset.

We used the fine-tuning strategy, as well as the stochastic gradient descent with momentum optimizer (SGDM) at their default values, dropout rate set at 0.5 and early-stopping to prevent over-fitting, and the learning rate at 0.0001. SGDM is used to accelerate gradients vectors in the correct directions, as we do not need to check all the training examples to have knowledge of the direction of the decreasing slope, thus leading to faster converging. Additionally, to balance the memory consumption and the performance of the network, we used a batch size of 32 to update the network weights more often, and trained them in 30 epochs.

To preprocess images further, all samples go through a heavy data augmentation process which includes flipping, 180^o random rotation, rescaling factor between 70% and 130%, and horizontal and vertical translations between -20 and $+20$ pixels. The networks were trained using the ONNX package for Matlab® on a computing node of the CeDRI cluster equipped with a NVIDIA A100 GPU.

We used 5-Fold Cross-Validation, where in each set the images were split on five folds, allowing the networks to be independently trained and validated on different sets. Since each testing set is built with images not seen by the training model, this allows us to anticipate the network behaviour against new images.

4 Results and Discussion

All the proposed approaches available in the literature that applied deep learning techniques to classify the *Kaggle* EyePACS, APTOS or DDR datasets used the imbalanced version of the datasets. Table 3 summarizes the accuracy for each class and the overall accuracy (OA) for the three DR datasets. In the approaches that used the DDR dataset we have removed grade 5 samples.

When using imbalanced datasets, the accuracy tends to bias towards majority classes [22]. Analysing Table 1 and Table 3 we can see that these results correlate well with the number of samples used in each class. If we use a balanced dataset with the samples in the classes equally distributed then the DLN can learn the features properly, but in the situation of unequal distribution, DLN over-fits

Table 3. DR grading classification on the test set for *Kaggle* EyePACS, APTOS and DDR datasets in terms of accuracy.

Dataset	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4	OA
EyePACS [21][1]	0.950	0.000	0.233	0.085	0.443	0.738
EyePACS [22][2]	0.973	0.546	0.509	0.617	0.596	0.648
APTOS [10]	0.986	0.476	0.727	0.250	0.083	0.766
APTOS [1]	0.978	0.729	0.860	0.000	0.644	0.842
DDR [17]	0.945	0.048	0.646	0.127	0.582	0.751
DDR [1]	0.988	0.254	0.926	0.234	0.648	0.890

[1] Pratt et al. [21] have used only a subset of 5000 samples from EyePACS.
[2] Qummar et al. [22] have used a balanced EyePACS using data augmentation.

for the highly sampled class. The OA is obtained as a weighted average of the accuracy grades and, as the weight of grade 0 is very high, the importance of an accuracy of 0.0 for grade 1 in the work [21] does not change the OA value (e.g., in the EyePACS dataset an accuracy of 0.95 for grade 0 contributes with a minimum value of 0.70 for the overall accuracy).

In our experiments we used a balanced version of the DRR dataset, by completing the grade sets with samples from several DR datasets. This approach has two advantages over other proposed balanced datasets: first, using more than one dataset allows to incorporate variability into the DLN model, which increases the generalization by reducing sensitivity to the image acquisition system; secondly, it uses different real images to complete the grade sets. Other authors have proposed data augmentation to balance the datasets [1,22]. In spite of that, they used a small number of real samples, which will produce low variability in the grade set, specially in the mild DR set.

In our judgement, the results obtained with imbalanced datasets are heavily biased and, as far as we know, we are the first to study the real accuracy obtained by DLN when using a dataset of balanced grade sets by using real images obtained from different publicly available datasets.

Table 4 summarizes the accuracy results for each class obtained from ten state-of-the-art DLNs, trained with 5-Fold cross-validation. These DLNs were trained and tested over the balanced dataset described in Table 2 with 15665 samples for the training set, 6265 samples for the validation set and 9400 samples for the test dataset, equally distributed over all five DR grades.

The experimental results demonstrate that the results obtained with the DL networks do not differ substantially in the overall accuracy, where the VGG16 network shows the highest overall accuracy, but a comparatively low accuracy on grades 2 and 4. The high accuracy for grade 0 (normal retina) detection suggests that we can use these approaches to rapidly make a binary classification, in normal and DR retina. All these networks exhibit poor performance on sam-

Table 4. Classification results (accuracy) of individual DLN classifiers on the test set for the different DR grading classes considered.

DLN	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4	OA
VGG16	0.991	0.693	0.359	0.619	0.429	0.618
VGG19	0.952	0.803	0.244	0.815	0.227	0.608
ResNet50	0.949	0.771	0.317	0.560	0.334	0.586
ResNet101	0.956	0.753	0.323	0.613	0.256	0.580
Inception-V3	0.960	0.732	0.292	0.703	0.289	0.595
Incep.-ResNet	0.970	0.750	0.262	0.673	0.359	0.603
Xception	0.952	0.697	0.261	0.668	0.323	0.580
DenseNet201	0.960	0.804	0.366	0.568	0.285	0.597
DarkNet53	0.940	0.805	0.452	0.539	0.301	0.607
EfficientNetB0	0.967	0.731	0.234	0.702	0.284	0.584

ples with moderate and proliferative DR, suggesting that these models are weak in learning these two types of samples. Furthermore, swelling and distortion of blood vessels in moderate DR are difficult to identify, while samples of proliferative DR are easily misclassified as moderate or severe DR, namely when there are no severe hemorrhages.

The confusion matrix in Fig. 4 shows the misclassifications produced by the VGG16 model when applied to a DR grading prediction task. Note that the test set is composed by 9400 samples, equally distributed with 1880 samples for each DR grade. In the confusion matrix, we can see that most of the moderate DR type samples are predicted as normal.

5 Conclusion

Diabetic retinopathy is a complication of diabetes that is mainly caused by the damage of the blood vessels located in the retina. Computer-aided diagnosis approaches are needed to allow an early detection and treatment.

In this paper, we assessed the state-of-the-art deep learning models for DR grading. In our experiments we balanced the DDR dataset using images from the EyePACS, APTOS, Messidor-2 and IDRiD datasets, so that all classes are represented equally. This new balanced dataset aims to reduce biased classification presented in approaches that used imbalanced datasets.

We trained and tested ten individual DLN models, with 5-fold cross-validation. Our results confirm that most of the published approaches that used imbalanced datasets are biased by the cardinality in each grade set, as the overall accuracy benefits the largest grade sets.

In the future, we plan to design a blended framework that combines DR grading from the DLNs and combines their predictions in a final score. This ap-

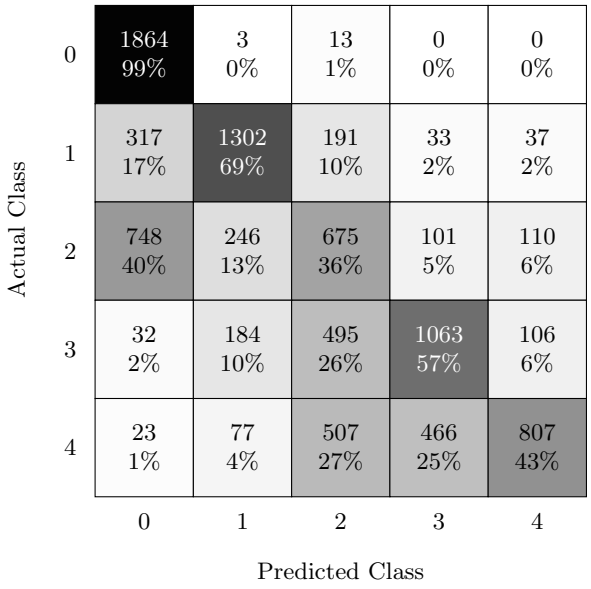


Fig. 4. Confusion matrix of the VGG16 network for the DR grading task.

proach has the advantage of training DLNs in different resolutions and increases the robustness by reducing possible over-fitting in individual models.

Acknowledgments

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI (UIDB/05757/2020 and UIDP/05757/2020) and SusTEC (LA/P/0007/2021).

References

1. Alyoubi, W.L., Abulkhair, M.F., Shalash, W.M.: Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors **21**(11), 3704 (2021)
2. Alyoubi, W.L., Shalash, W.M., Abulkhair, M.F.: Diabetic retinopathy detection through deep learning techniques: a review. Informatics in Medicine Unlocked **20**, 100377 (2020)
3. Asia Pacific Tele-Ophthalmology Society: Aptos 2019 blindness detection (2019), https://www.kaggle.com/competitions/aptos2019-blindness-detection, accessed on 4 April 2022.
4. Baker, N., Lu, H., Erlikhman, G., Kellman, P.J.: Local features and global shape information in object classification by deep convolutional neural networks. Vision Research **172**, 46–61 (2020)

5. Bhatia, K., Arora, S., Tomar, R.: Diagnosis of diabetic retinopathy using machine learning classification algorithm. In: 2016 2nd International Conference on Next Generation Computing Technologies. pp. 347–351 (2016)
6. Bodapati, J.D., Naralasetti, V., Shareef, S.N., Hakak, S., Bilal, M., Maddikunta, P.K.R., Jo, O.: Blended multi-modal deep ConvNet features for diabetic retinopathy severity prediction. Electronics **9**(6), 914 (2020)
7. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1800–1807 (2017)
8. Dai, L., Wu, L., Li, H., cai, C., Wu, Q., Kong, H., et al., R.L.: A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature Communications **12**(1), 3242 (2021)
9. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., Klein, J.: Feedback on a publicly distributed image database: The Messidor database. Image Analysis & Stereology **33**(3) (2014)
10. Dekhil, O., Naglah, A., Shaban, M., Ghazal, M., Taher, F., Elbaz, A.: Deep learning based method for computer aided diagnosis of diabetic retinopathy. In: 2019 IEEE International Conference on Imaging Systems and Techniques (IST). pp. 1–4 (2019)
11. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
12. EyePACS: Diabetic retinopathy detection (2015), https://www.kaggle.com/c/diabetic-retinopathy-detection, accessed on 4 April 2022.
13. Gargeya, R., Leng, T.: Automated identification of diabetic retinopathy using deep learning. Ophthalmology **124**(7), 962–969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017)
17. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
18. Lu, J., Xu, Y., Chen, M., Luo, Y.: A coarse-to-fine fully convolutional neural network for fundus vessel segmentation. Symmetry **10**(11), 607 (2018)
19. Majumder, S., Kehtarnavaz, N.: Multitasking deep learning model for detection of five stages of diabetic retinopathy. IEEE Access **9**, 123220–123230 (2021)
20. Porwal, P., Pachade, S., Kokare, M., et al.: IDRiD: Diabetic retinopathy - segmentation and grading challenge. Medical Image Analysis **59**, 101561 (2020)
21. Pratt, H., Coenen, F., Broadbent, D.M., Harding, S.P., Zheng, Y.: Convolutional neural networks for diabetic retinopathy. Procedia Comp. Sci. **90**, 200–205 (2016)
22. Qummar, S., Khan, F.G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z.U., Khan, I.A., Jadoon, W.: A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access **7**, 150530–150539 (2019)
23. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. ArXiv p. 1804.02767 (2018)

24. Rocha, D.A., Ferreira, F., Peixoto, Z.: Diabetic retinopathy classification using vgg16 neural network. Research on Biomedical Engineering **38**, 761–772 (2022)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Int. Conference on Learning Representations. pp. 1–14 (2015)
26. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inceptionresnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. pp. 4278–4284 (2017)
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
28. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 6105–6114 (2019)
29. Taylor, R., Batey, D.: Handbook of retinal screening in diabetes: diagnosis and management, second ed. Wiley-Blackwell (2012)
30. Teo, Z., Tham, Y., Yu, M., Chee, M., Rim, T., Cheung, N., Bikbov, M., Wang, Y., Tang, Y., et al.: Diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology **128**(11), 1580–1591 (2021)
31. Tsiknakis, N., Theodoropoulos, D., Manikis, G., Ktistakis, E., Boutsora, O., Berto, A., Scarpa, F., Scarpa, A., Fotiadis, D.I., Marias, K.: Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. Computers in Biology and Medicine **135**, 104599 (2021)
32. Wan, S., Liang, Y., Zhang, Y.: Deep convolutional neural networks for diabetic retinopathy detection by image classification. Computers & Electrical Engineering **72**, 274–282 (2018)
33. Wilkinson, C., Ferris, F., Klein, R., Lee, P., Agardh, C., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**(9), 1677–1682 (2003)
34. World Health Organization: World report on vision - Licence: CC BY-NC-SA 3.0 IGO (2019), https://www.who.int/publications/i/item/9789241516570, accessed on 26 April 2022.
35. Zago, G.T., Andreão, R.V., Dorizzi, B., Teatini Salles, E.O.: Diabetic retinopathy detection using red lesion localization and convolutional neural networks. Computers in Biology and Medicine **116**, 103537 (2020)