Deep Residual Transfer Learning for Automatic Diabetic Retinopathy Grading

Francisco J. Martínez Murcia[b,c], Andrés Ortiz[a,c,*], Javier Ramírez[b,c], Juan M. Górriz[b,c], Ricardo Cruz[a]

[a]*Department of Communications Engineering, University of Málaga*
[b]*Department of Signal Theory, Communications and Networking. University of Granada*
[c]*Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI)*

Abstract

Evaluation and diagnosis of retina pathology is usually made via the analysis of different image modalities that allow to explore its structure. The most popular retina image method is retinography, a technique that displays the fundus of the eye, including the retina and other structures. Retinography is the most common imaging method to diagnose retina diseases such as Diabetic Retinopathy (DB) or Macular Edema (ME). However, retinography evaluation to score the image according to the disease grade presents difficulties due to differences in contrast, brightness and the presence of artifacts. Therefore, it is mainly done via manual analysis; a time consuming task that requires a trained clinician to examine and evaluate the images. In this paper, we present a computer aided diagnosis tool that takes advantage of the performance provided by deep learning architectures for image analysis. Our proposal is based on a deep residual convolutional neural network for extracting discriminatory features with no prior complex image transformations to enhance the image quality or to highlight specific structures. Moreover, we used the transfer learning paradigm to reuse layers from deep neural networks previously trained on the ImageNet dataset, under the hypothesis that first layers capture abstract features than can be reused for different problems. Experiments using different convolutional architectures have been carried out and their performance has been evaluated on the MESSIDOR database using cross-validation. Best results were found using a ResNet50-based architecture, showing an AUC of 0.93 for grades 0+1, AUC of 0.81 for grade 2 and AUC of 0.92 for grade 3 labelling, as well as AUCs higher than 0.97 when considering a binary classification problem (grades 0 vs 3).

Keywords: Deep Learning, Residual Learning, Transfer Learning, Convolutional neural network, Retinography, Diabetic Retinopathy

1. Introduction

Diabetic Retinopathy (DR) is an eye disease consisting on retina damage due to diabetes mellitus, a chronic degenerative disease with a worldwide prevalence of 2-6% [1]. An early diagnosis and treatment diminishes the risk

*Corresponding Author. Tel: +34 952133353

Email addresses: fjmm@ic.uma.es (Francisco J. Martínez Murcia), aortiz@ic.uma.es (Andrés Ortiz), javierrp@ugr.es (Javier Ramírez), gorriz@ugr.es (Juan M. Górriz), rcruz@ic.uma.es (Ricardo Cruz)

of disease progression, which is crucial to prevent permanent damage to the vision system. With an increasing number of diabetic patients, DB has become the leading cause of blindness in the working-age population of the developed world, affecting over 93 million people worldwide. DR diagnosis is closely related to the stage of the disease, and it is usually performed by finding abnormalities in eye fundus (retina).

The retina is a layered tissue that constitutes an essential part of the human vision system. It is a sensorial membrane that recover the inner part of the back of the eye. Different parts of the retina contain specialized cells called photoreceptors, which are responsible for the extraction of different features that eventually allows to distinguish not only objects but also motion. Signals produced by photoreceptors are driven to the visual brain cortex for further and complex processing such as texture detection, object recognition, etc. As a consequence, the health of the retina directly affects the quality of life of the population. In this way, an accurate diagnosis of retinal pathologies allows an early treatment that maximizes the probability of success, since most retinal pathologies involve a degenerative process.

DR can be classified according to four severity retinopathy grades (RG), assessed by the presence of abnormalities such as microaneusysms, intraretinal hemorrhages and neovascularization found at different parts of the retina [2, 3]. The most common technique for revealing these abnormalities in the eye fundus and then, for screening retinopathy in diabetic people is retinography, which provides an image of the eye fundus. Depending on the type of retinograph it may need the patient's pupils to be dilated (midriatic) or nor (non-midriatic). Optical coherence tomography (OCT) is a more recent imaging technique that uses a light beam to scan the inner retina, producing cross-sectional images of the different layers in the retina. But despite the technique used, the identification of signs to grade the disease stage still requires a trained, expert clinician and a considerable time of manual processing.

All the same, the diagnosis of retinal pathologies is not always evident from retinal images, specially in the early stages of the disease. Nevertheless, current Computer-aided image processing tools have demonstrated its ability in clearly identifying abnormal patterns linked to a disease [4, 5, 6]. Since blood vessel segmentation is commonly considered a first step in the construction of computer-aided diagnostic (CAD) tools, a number of methods have been developed in recent years for the removal of blood vessels from retinography using classic image processing and automatic learning methods [7, 8].

The success of Deep Learning [9] in image classification tasks has revolutionized applications in many fields such as medical imaging. Particularly, Convolutional Neural Networks (CNNs) have largely outperformed previous methods based on statistical learning in many imaging applications. As a result, different Deep Learning methods for retinal image analysis and classification have flourished in the last years, e.g., blood vessel segmentation using different architectures that include convolutional layers [10, 11] (it is worth noting that segmentation is in fact a classification task that aims to split the image into blood vessels and non blood vessels). Most current segmentation approaches, e.g. the U-NET architecture [10], need a complex preprocessing stage that may include filtering and morphological transformations as well as histogram equalization over the any of the RGB channels to enhance images in order to highlight the blood vessel structure. At the same time, the CNN architectures are simplified to use only one input channel instead of all colour channels (RGB).

Previous approaches aim to reveal vessel structures from the image. These can extract statistical features that characterize the vessel distribution at different parts of the image, which will be eventually used for classification. However, these methods are not specifically developed to extract discriminative information from the image (among classes), but representative features for segmentation, in order to facilitate their interpretability by expert clinicians. Advanced image processing techniques and its synergy with artificial intelligence methods make possible the development of classification approaches taking into account all the information contained in the eye fundus image and not only the blood vessel distribution in the retina.

In this work, we propose the use of a deep CNN architecture for automatic DR grading with no prior complex image transformations to enhance the image quality or to highlight specific structures. The use of residual networks allows to overcome this and further difficulties during the training process, such as overfitting and vanishing gradients. Additionally, we used transfer learning for building our residual networks by reusing the topmost layers of previously trained CNNs with general content images from the ImageNet database [12] not specifically related to retina.

The rest of the paper is organized as follows. Section 2 depicts the related work regarding automatic retinography processing and classification, along with diabetic retinopathy grading. Then, the database and methods are described in the Materials and Methods section. At Section 4 we present the results obtained and the evaluation of the proposed approaches, along with the discussion provided in Section 5. Finally, the main conclusions of this work are drawn in Section 6.

2. Related Work

Early identification of retinal pathologies is crucial to stopping or at least delaying the damage to the visual system caused by DR. The task, which requires both expertise and a considerable amount of time, is frequently performed via visual inspection of retinopathy images. At this point, image preprocessing techniques may facilitate the task, by extracting regions of interest, isolating the structures involved in the diagnosis and grading and removing image artifacts due to the acquisition process. Moreover, the use of statistical and learning-based methods to extract discriminative features constitutes an effective way to develop CAD tools to leverage the detection accuracy of DR, especially in the early stages of the disease.

Nevertheless, classification of non-midriatic eye fundus images retinography images poses a challenge for the screening and diagnosis of DR [13] (for instance, non-midriatic cameras are more prone to generate under and over-exposure artifacts). Thus, many works are focused on segmentation of retinal images in order to extract the blood vessel structure or to recognize clinical features such as microaneurysms, hemorrhages, hard exudates and soft exudates [14]. Other works extract statistical features and then use an statistical classifier, such as [15], which uses common invariant features in image retrieval such as local binary patterns (LBP) [16], SIFT features (Scale Invariant Feature Transform) [17] and LDP (Local Directional Patterns) features [18]. Then, the new features can be classified by a Bag-of-Words model. These features are commonly used to diagnose DR and to grade it according to a severity scale [2].

Most recent works in the field make use of deep learning architectures in segmentation and classification. It has demonstrated flexibility, adaptability and unprecedented accuracy in image classification applications, which has motivated its adoption in many disciplines ranging from speech[19] to pattern recognition [20] or drug discovery [21]. It has also been applied in the analysis and classification of medical images [5, 22, 23] with great success. But, most importantly, Deep Learning techniques for image classification show their effectiveness under very complex conditions (i.e. differences in orientation, scaling, occlusion, etc.), and are very robust against noise and artifacts due to their ability to extract discriminant patterns. Thus, it provides an appropriate arena to process and classify eye fundus images, since they are not always acquired under the same conditions and they often contain artifacts [14]. In this regard, [24] proposes a segmentation method based on ensemble learning to combine Random forests classifier with CNNs. In [25] a method is proposed for blood vessel segmentation using CNNs, and then, classification is performed by means of the segmented images, extracting discriminant patterns from the blood vessel structure. A similar idea is used in [26], which proposes the use of an hybrid method to extract structure descriptors based on Gabor filters and textural features, and the use of CNNs to classify the image features. Along the same line, [27] shows an automated method for localizing and discerning lesions in retinal images using different types of pre-trained CNNs.

Many recent works address the problem of retinopathy classification. This aims to differentiate among different retinal diseases that are expressed in the image by different abnormalities. In [28], an VGG19 network [29] is trained using data augmentation on the STARE database to differentiate among 10 retinal diseases. Moreover, experiments using the CNN only to extract features to be eventually classified by an SVM are also carried out. This is the case of [30], which shows the classification performance provided by different AlexNet and GoogleNet CNN architectures directly trained using a publicly available Kaggle dataset of 35000 retinal images. In [31], a shallow CNN architecture composed of 13 layers is proposed allowing similar performance than deeper structures in binary classification (controls vs. severe DR). Similarly, [32] proposes a deep CNN that shows AUC up to 0.9 for classifying between controls and severe (stage 4) DR. Similarly, [33] presents a CNN based on the AlexNet architecture [34], composed of 4 convolutional layers for extracting features and 3 fully connected layers for classification. In [35], a CNN classifier is built to generate easily interpretative visual maps to help the ophthalmologist in the diagnosis of the DR. Finally, in [36], a network architecture using residual blocks is proposed for automatic classification of retinopathy images, achieving accuracy up to 81%. In all those works, DR grading is addressed as a binary classification problem, to distinguish between controls and different levels of DR severity, usually the most advanced state.

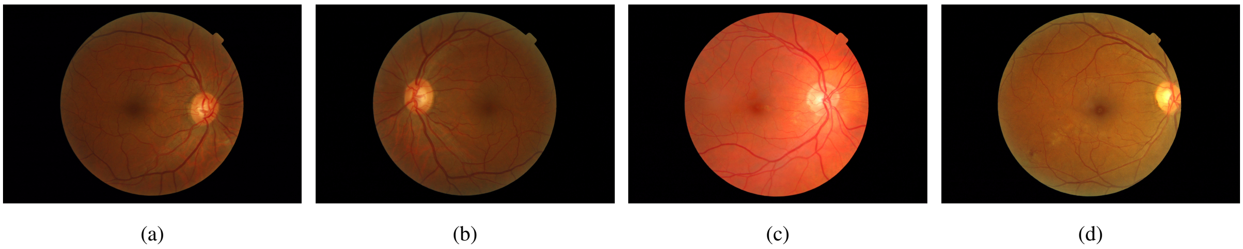


Figure 1: Example of retina images for a) RG=0, b) RG=1, c) RG=2 and d) RG=3. Note the differences in contrast, brightness and illumination of different acquisitions. Moreover, differences between classes are subtle at first sight, especially between grades 0 and 1.

3. Materials and Methods

3.1. Database

The classification experiments shown in this work has been carried out using the MESSIDOR database [37]. MESSIDOR project (Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology) was born within the scope of Diabetic Retinopathy, to compare and evaluate computer-assisted diagnosis methods and algorithm for DR.

The database is composed of 1200 eye fundus color images acquired by 3 ophthalmologic departments (at Paris, Etienne and Brest) using a color video 3CCD camera mounted on a Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view. Images were captured using 8 bits per color plane (R, G, B) at 1440*960, 2240*1488 or 2304*1536 pixels. 800 images were acquired with pupil dilation (one drop of Tropicamide at 0.5%) and 400 without dilation. Two diagnoses have been provided by the medical experts for each image, including Retinopathy grade and Risk of macular edema, according to the number of microaneusysms (μA), intraretinal hemorrhages (H) and neovascularization (NV): 0- No abnormalities (normal image); 1- Mild Non-Proliferative Diabetic Retinopathy ($0 < \mu A \leq 5$, $H = 0$); 2- Non-Proliferative Diabetic Retinopathy ($5 < \mu A < 15$ or $0 < H < 5$, and no NV); 3- Proliferative Diabetic Retinopathy (either $\mu A \geq 15$ or $H \geq 5$ or NV). Examples of retina images corresponding to the aforementioned labels are shown in Figure 1.

The rectangular images were cropped down to a square containing the retina circle. Should that be necessary, rescaling was used to ensure an image size of 900x900 pixels. The image intensity was independently scaled to the range $[0, 1]$ in each channel to facilitate convergence of the network training, with no further equalisation or image enhancing.

3.2. Residual Neural Networks

Very-deep convolutional neural networks such as VGGs [29] have been the standard in feature extraction and classification challenges for many years. However, this deepness comes at a price, and the vanishing gradient was one of their biggest drawbacks. In order to overcome this, He et al. [38, 39] proposed the residual networks, or ResNet, as a solution to the vanishing gradient in very deep convolutional networks, and have afterwards become the de-facto standard in computer vision applications [9].

Residual neural networks are composed by several layers, each of which is in turn composed by a series of residual blocks. The residual block is the most basic building block of a residual network. It processes an input signal $\mathbf{y}_{i-1}$ like any other convolutional network, via convolution, normalization and activation layers. But it includes a “skip” connection that combines the output of all those layers with the input signal itself, in the form:

$$\mathbf{y}_i = f_i(\mathbf{y}_{i-1}, \mathbf{W}_i) + \mathbf{y}_{i-1} \tag{1}$$

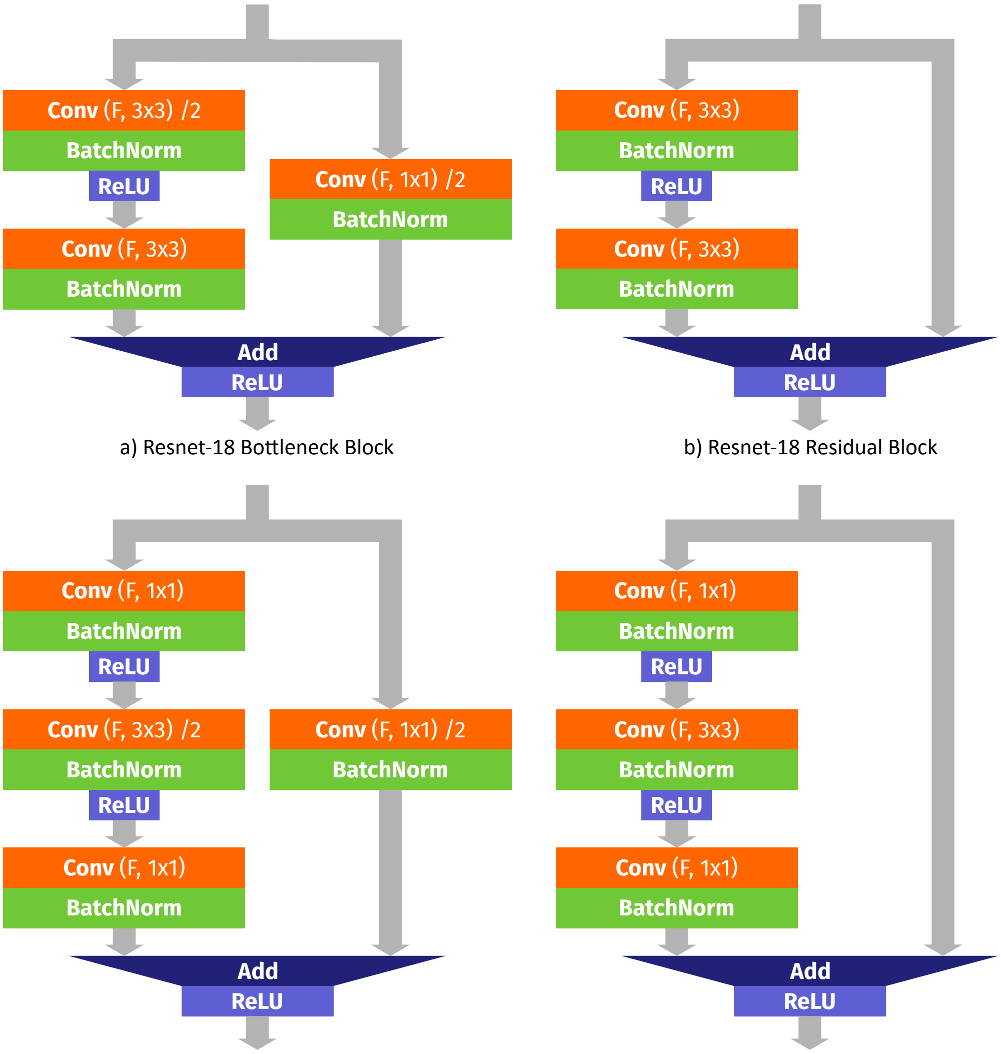
where $\mathbf{y}_i$ is the output of the residual block, $\mathbf{W}_i$ is the set of learnable parameters of the block and $f(\cdot)$ is a trainable non-linear mapping, which is usually a succession of convolution, normalization and activation layers. The skipped link is what helps tackling the vanishing gradient, by helping to modulate the activation of the previous layer during training, and amplifying the previously-skipped layer. Figure 2 shows different options for residual blocks.

There is also a special residual block intended to reduce dimensionality in the input space: the *bottleneck* block. In this block, the skipping path (the input signal not modified by $f_i(\cdot)$) passes through a convolution layer with stride and normalization. Then, it is combined with the output of the regular path, which has as well included strided convolutions in any of its layers, so that the sizes match. A schema of these building blocks can be checked at Figure 2.

Note that there are two different versions for the residual and the blottleneck blocks. The first, shallower blocks were the first introduced in [39], which were used in building a 18 and 36-layer residual network. Later in the paper, they introduce 50, 101 and deeper models in which they replaced the 2-layer initial blocks by 3-layer blocks that increased dimensions. In this work, we will test two different architectures: the **ResNet-18** and **ResNet-50**, which combine fewer parameter than other very-deep networks (the 101 or 152 -layer versions) with low error in the ImageNet challenge [12].

3.3. Transfer Learning

As aforementioned, Deep Learning is a powerful learning paradigm that allows to build very complex models containing thousands or even, millions of parameters. However, the high number of parameters makes Deep Learning models very prone to overfit the data. This limits the generalization capabilities, making it necessary to adopt different strategies to improve the performance of the model when facing new and previously unseen data. In this way, regularization methods are commonly adopted to fight against overfitting, forcing the model to correctly classify training samples when the learning process is discouraged by adding noise to the parameters or constraining their values. Nevertheless, the first step to reduce the generalization error is to enlarge the number of training samples for a best representation of the data manifold, but, unfortunately, it is usually difficult or even impossible to get new data samples, specially in biomedical applications where data acquisition is expensive and requires time. At this point, it is possible to take advantage of the data abstractions generated in the different layers of a Deep network: the upper the layer, the more abstract is the data stored on it, to the point of being independent of the training samples when the



a) Resnet-18 Bottleneck Block

b) Resnet-18 Residual Block

c) Resnet-50 Bottleneck Block

d) Resnet-50 Residual Block

Figure 2: Schema of the building blocks of a residual network: the bottleneck block that usually downsamples the input by a factor of 2 using convolutions with stride, indicated with "/2" for ResNet-18 (a) and ResNet-50 (c), and the residual block for ResNet-18 (b) and ResNet-50 (d). Size of convolution kernels is indicated as, e.g., 3×3, and the letter F stands for the number of filters, which varies from one layer to another.

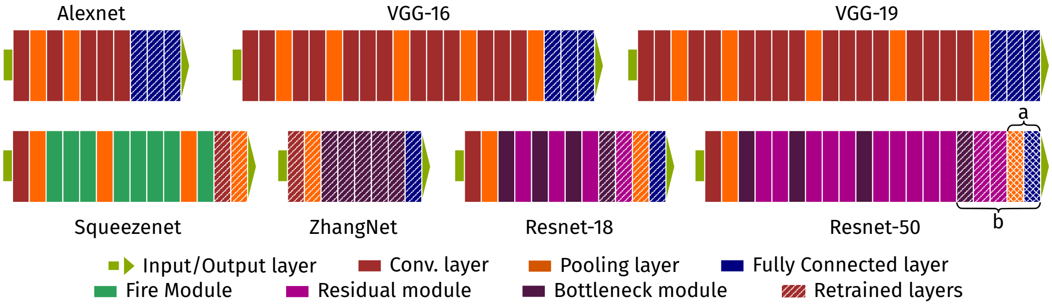


Figure 3: Schema of the architectures for the different network topologies and the retrained layers.

model is generated using a vast number of training samples. Hence, a model that has been trained using a large number of images has probably stored enough abstract information to be reused in other problems for feature extraction. This is called Transfer Learning (TL) [40]. TL consist on the reuse of a number of layers (usually the upper most) of an already general trained model and adapt the lower layers to a specific problem. In the case of convolutional neural networks, it includes the modification of the first convolutional layers (to adapt to the current image size), and the inclusion of a specific fully connected block to implement the classifier according to the number of classes being classified.

The training process using the TL paradigm is usually performed in two stages. First, the layers whose weights will be reused, are fixed to avoid their weights updating via backpropagation. Then, Adadelta [cite] (lr=1.0, ρ=0.9) is used to update the parameters of the selected layers, and early stopping [citeEarlyStopping] is used to set the weights to the minimum loss iteration. Early stopping is described here as the iteration with minimum loss over a 20-epoch training. Finally, in a further finetuning stage, we use Stochastic Gradient Descent (SGD) with low learning rate (lr=0.001) over 50 epochs and early stopping for the final trained model.

In this work, different deep convolutional networks have been used along with TL, where the models have been trained using the Imagenet database [12], a very large database designed for use in visual object recognition task, composed of more than 14 million hand-annotated images. For more detail on the retrained layers, see Figure 3. All parameters on the highlighted layers are retrained in the finetuning step. Note that in ResNet-50, two re-training cases are proposed: in ResNet-50a only the last fully-connected layers are re-trained, but for ResNet-50b (and ResNet-18) the last residual layer has also been re-trained with the new retina data.

3.4. Other Architectures

In order to compare our ResNet-based system with the state of the art, we have trained and tested under the same conditions four additional, well-known, deep learning models: AlexNet [34], SqueezeNet [41], VGGnet-16 and VGGnet-19 [29], and an additional residual network, ZhangNet [36].

AlexNet was perhaps the major breakthrough in Convolutional Neural Networks [34]. It achieved unprecedented accuracy in the ImageNet challenge [12], and paved the way to the deep learning revolution. It is a simple but rather effective architecture, that has been used and adapted in many works [42]. SqueezeNet [41], for its part, was an all-convolutional network intended to achieve performances smilar to AlexNet but significantly reducing the number of parameters. We also will use two versions of the high-scoring VGG [29] networks, with 16 and 18 layer respectively, which were the standard in image analysis for a long time, and a fusion-based residual network for retinopathy classification, ZhangNet, presented in a conference [36]. Since no pretrained model for ZhangNet was available, it was fully trained on our dataset. All the models have been coded using the PyTorch framework.

The details of these architectures are shown at Figure 3, where the retrained layers are also specified. The individual layer parameters (number of layers, filters, size of the filters, etc) are kept as in the original papers.

4. Results

4.1. Evaluation

In this work, a transfer learning scheme for the automatic retinopathy grading based on ResNet is proposed. We also analyze four additional widely-used models (AlexNet, ResNet, VGG-16 and VGG-19), all tested within a 5-fold stratified cross-validation. During training of all models, data augmentation via random horizontal flipping (with $p = 0.5$) was applied. Evaluation parameters such as global correct rate (CR), sensitivity and specificity (Sens. and Spec., only for binary classification) and the Receiver Operating Characteristic (ROC) curve, as well as the area under the ROC (AUC) were computed per class. Retinopathy grading is not just a usual multiclass classification problem, but one in which the different classes (see Section 3.1) indicate a higher degree of a measure, which is frequently known as ordinal regression. Therefore, measures typically related to regression, such as the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) between the output and the original labels can be used to obtain a performance measure that penalize non-adjacent mistakes (e.g. labelling a 0-grade as 3-grade). Additionally, the full normalized confusion matrix, a matrix containing the proportion of samples of any class labelled as any of the possible outputs, is also provided.

These results are provided for three different setups. In the first approach, we present the most challenging scenario: a full multi-class grading using the different models on the four different labels. Then, we test an aggregation of the 0 and 1 labels, similar to the ones proposed in [35]. Finally, we also provide an analysis of the performance of a binary classification scenario covering grades 0 (control) vs 3.

4.2. Multi-class Model Performance

In this section we provide the results for the automatic 4-classes grading system. The results for all models is provided in the form of confusion matrices at Figure 4.

From Figure 4, it is easy to assess AlexNet fails completely when classifying the images into the 0, 1 and 2 classes, only achieving a reasonable performance in class 3. Both the VGGs and the ResNet have a tendency to grade

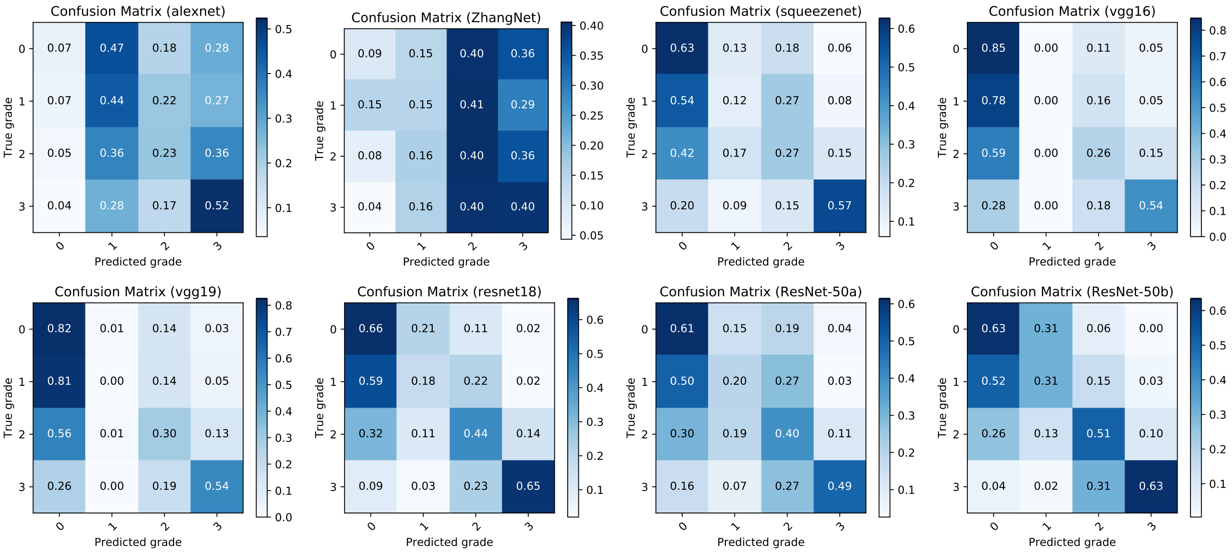


Figure 4: Normalized confusion matrices for the eight tested models.

all dubious samples as 0, making their confusion matrices far from the expected diagonal matrix. In this context, the best performing models are the ResNet-based (ResNet18 and ResNet50), that achieve a higher classification accuracy of grade-3 retina images, in contrast to other approaches, at the same time that keep or even improve the performance of classes 0 and 3.

				AUC (grade)			
model	CR	MAE	MSE	0	1	2	3
AlexNet:	0.246 [0.06]	1.211 [0.19]	2.394 [0.76]	0.596	0.540	0.533	0.696
ZhangNet:	0.229 [0.04]	1.386 [0.26]	2.964 [1.19]	0.491	0.522	0.527	0.547
SqueezeNet:	0.476 [0.08]	0.862 [0.11]	1.676 [0.25]	0.711	0.591	0.617	0.809
VGG-16:	0.552 [0.05]	0.788 [0.14]	1.632 [0.43]	0.699	0.584	0.595	0.819
VGG-19:	0.551 [0.05]	0.775 [0.12]	1.569 [0.35]	0.697	0.597	0.595	0.817
ResNet-18:	0.550 [0.11]	0.631 [0.13]	1.051 [0.23]	0.776	0.610	0.738	0.912
ResNet-50a:	0.490 [0.10]	0.786 [0.13]	1.443 [0.25]	0.724	0.557	0.663	0.893
ResNet-50b:	**0.564 [0.13]**	**0.544 [0.14]**	**0.782 [0.15]**	**0.807**	**0.643**	**0.786**	**0.936**

Table 1: Performance metrics of the different models analyzed.

This can be also reflected at the performance metrics displayed at Table 1. In this case, we can actually look at the average Correct Rate (and its standard deviation, in brackets) with regards to the model used. With respect to CR, our ResNet models look very similar to the VGGs. However, since MAE, MSE and AUC for each label is also provided, it gives a complete quantitative idea of the performance of the different strategies. ResNet reveal themselves as better in distinguishing classes 1 and 2, as it can be seen from their AUCs. But in addition to this, the lowest MAE and MSE values indicate that the classification errors are fundamentally distributed in adjacent classes (as we could also check at the confusion matrices), making them the best models in for this automatic grading system. From ResNet-50 a and b, we can infer that there is a significant difference between retraining or not the last residual layer, therefore the features learnt by this layer should be of fundamental importance for detecting retinopathy-related lesions.

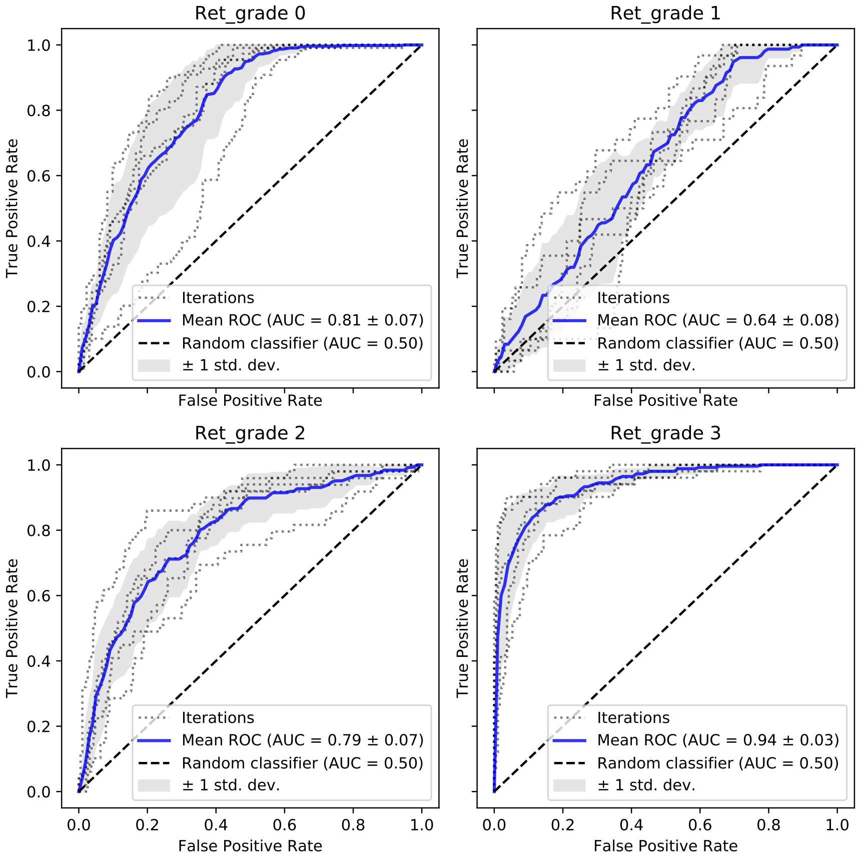


Figure 5: ROC curves for the ResNet-50b model. Note the confusion in grading labels 0 and 1.

This is also confirmed when plotting the ROC curves for the cross-validation iterations of the ResNet-50b model (Figure 5). There, the biggest confusion is obtained for grade 1, which is often confused with grade 0 (as we saw previously in Figure 4), but keeping a good performance for grade 2 and excellent in grade 3. Together with the distribution of errors, this performance demonstrates that the automatic retinopathy grading system based on ResNet could be most optimal for further use.

4.3. Aggregated Grading: 3-class Scenario

In addition to the results exposed in Table 1, we also wanted to check whether combining 0 and 1 in one class would improve grading performance, in what defines a 3-class problem. Under these conditions, the overall accuracy of the models improve significantly, achieving AUCs higher than 0.8 in all classes (in fact, 0.93 for the combined 0+1 grades, and 0.918 for grade 3). A detailed comparison of the results of the analysed models can be found at Table 2

				AUC (grade)		
model	CR	MAE	MSE	0+1	2	3
AlexNet:	0.373 [0.08]	0.752 [0.12]	1.002 [0.23]	0.679	0.489	0.682
ZhangNet:	0.460 [0.21]	0.755 [0.34]	1.186 [0.70]	0.469	0.452	0.512
SqueezeNet:	0.611 [0.06]	0.519 [0.12]	0.779 [0.24]	0.759	0.629	0.807
VGG-16:	0.539 [0.08]	0.536 [0.13]	0.686 [0.22]	0.773	0.590	0.891
VGG-19:	0.613 [0.14]	0.481 [0.17]	0.667 [0.23]	0.753	0.607	0.835
ResNet-18:	0.731 [0.06]	0.293 [0.07]	0.341 [0.11]	0.894	0.730	0.901
ResNet-50a:	0.634 [0.05]	0.464 [0.09]	0.661 [0.18]	0.729	0.604	0.854
ResNet-50b:	**0.778 [0.05]**	**0.239 [0.05]**	**0.272 [0.06]**	**0.930**	**0.811**	**0.918**

Table 2: Results for the three-class grading problem.

In this case, the general accuracy improves significantly, but also all other measures such as the MAE and MSE and the AUC for class 2, in comparison to results of the 4-class problem. When comparing the confusion matrices of the ResNet-50b with the most basic one (AlexNet) and the more complex (VGG-19) in Fig. 6, we appreciate an enhanced grading of labels 2 and 3, as well as a reduced error between non-adjacent classes.

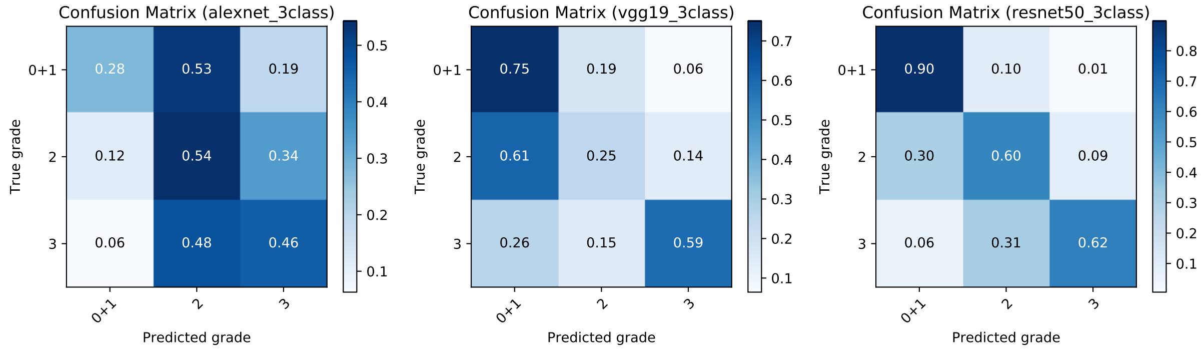


Figure 6: Comparison of the confusion matrix of the AlexNet, VGG-19 and ResNet-50b models within the 3-class scenario.

To provide a deeper insight into the accuracy of the ResNet-50b model, Figure 7 depicts the ROC curves for each of the cross-validation iterations and each grade. There it is far easier to appreciate the differences with respect to 0, 1 and 2 grading in the 4-class multivariate scenario (Fig. 5), especially with RG=2 and, above all, a significant decrease in the variance of these estimations, improving the consistence of the trained model.

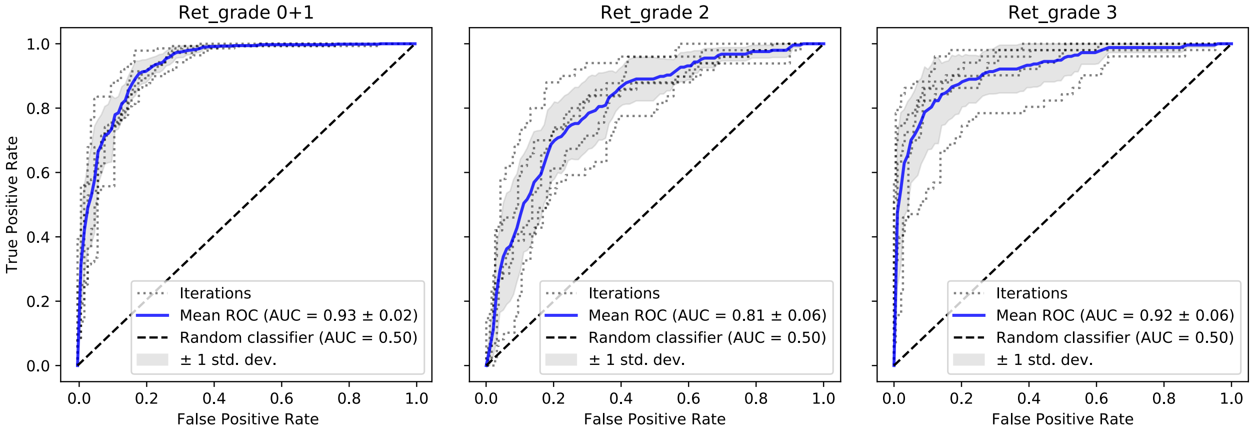


Figure 7: ROC curves for the ResNet50b 3-classes

4.4. Binary Classification: Grades 0 vs 3

Finally, in order to provide a scenario comparable to others in the literature [35, 30, 33, 28], we define a binary classification problem in which our system proves its ability to differentiate between grades 0 and 3. The results for this scenario are presented at Table 3.

				AUC (label)	
model	CR	Sens.	Spec.	0	3
AlexNet:	0.660 [0.15]	0.587 [0.18]	0.833 [0.18]	0.729	0.729
ZhangNet:	0.690 [0.15]	0.799 [0.00]	0.581 [0.15]	0.524	0.524
SqueezeNet:	0.791 [0.07]	0.737 [0.17]	0.838 [0.13]	0.834	0.833
VGG-16:	0.787 [0.07]	0.733 [0.16]	0.857 [0.14]	0.867	0.866
VGG-19:	0.792 [0.09]	0.735 [0.18]	0.861 [0.15]	0.871	0.870
ResNet-18:	0.926 [0.04]	0.975 [0.01]	0.912 [0.05]	0.956	0.955
ResNet-50a:	0.840 [0.06]	0.789 [0.12]	0.871 [0.10]	0.889	0.888
ResNet-50b:	**0.955 [0.02]**	**0.983 [0.02]**	**0.945 [0.03]**	**0.976**	**0.973**

Table 3: Results for the binary classification problem (grades 0 vs 3).

In this binary classification experiment the performances increase for all models. This is mainly due to a less complex problem in which classes are highly separable. Even AlexNet or ZhangNet, which achieved less than 0.4 of Correct Rate in other multi-class problems, achieves here almost 0.7. Again, this experiment confirms ResNet-50b as the better model for retinopathy grading in this paper, displaying outstanding performance (a CR of 0.963 [0.00]) when compared to other methods. This time, being a binary classification problem, the MAE and MSE measures are not a critical measure, and we provide Sensitivity and Specificity in their place.

5. Discussion

Experiments performed shown the superiority of convolutional networks including residual blocks with respect to non-residual architectures such as VGGnet. Deep neural networks usually contain a vast number of parameters. As a matter of fact, is it possible to construct very complex models of data. Nevertheless, the fact of having a huge number of parameters makes it difficult to learn simpler functions such as the identity function. This way, residual blocks, which add up the result of the previous layer to the output of the former bypassing it, ease the learning of simple functions. Consequently, residual networks avoid the degradation problem produced by a high number of layers and not related to overfitting but to the inability to model simpler data relationships (in fact, this is the curse of dimensionality problem [43]). On the other hand, they also mitigate the vanishing gradient problem that hinders the training process.

Transfer learning is, by definition, a very bio-inspired approach, since it learns to generalize from a huge diversity of examples, and therefore, a much smaller purpose-specific dataset is needed, yielding very generalizable results. And so is our approach, when compared to other works such as [28], which only used a relatively small retina database for training a VGG network, or [35] where the proposed network is trained using the EyePACS dataset (an eye fundus image database composed of more than 35000 images). The core idea behind transfer learning is that most abstract representations learned by a deep network can be used to generalize to very different data from which the network was trained. Thus, as explained in Sections 3.2 and 3.3, we used different layers of networks previously trained with the ImageNet dataset, composed of general content images but not related to retina.

We carried out experiments using different depth VGG and ResNet networks. As shown in Figure 4, ResNet-based models outperforms all other models not only in multiclass grading, but also in the 3-class classification and, above all, the binary classification problem. For the multiclass performance (Table 1), a MAE of 0.631 for ResNet-18 and 0.544 for ResNet-50b is achieved. Additionally, AUC values of 0.81, 0.64, 0.79 and 0.94 are obtained for grades 0, 1, 2 and 3, respectively as depicted in ROC curves shown in Figure 5. As expected, the performance obtained increases as DR becomes more evident in the images due to the appearance of more lesions in retina images. Conversely, VGG is able to correctly classify the most severe DR cases (grade 3) but it is unable to differentiate among mild ones (grades 1 and 2).

On the other hand, networks designed for the 4-class classification problem tend to confuse grades 0 (Controls) and 1 (mild DR). This was expected since lesions related to mild DR in retina images are very subtle and can often be confused with controls both by human and machine. As a result, grade 1 classification in multi-class mode reduces the overall performance, and especially degrades the classification accuracy of grade 2. This is demonstrated by aggregating grades 0 and 1 to the same group and then performing a multi-class classification experiment with 3 classes (grades 0+1, grade 2 and grade 3). In this case, MAE is drastically diminished to 0.239 for the ResNet-50b (see table 2 and ROC curves in Figure 7) and correct classification rate is improved for adjacent and non-adjacent classes, achieving AUC values of 0.930, 0.881 and 0.981 for grades 0+1, 2 and 3, respectively. Note that AUC for

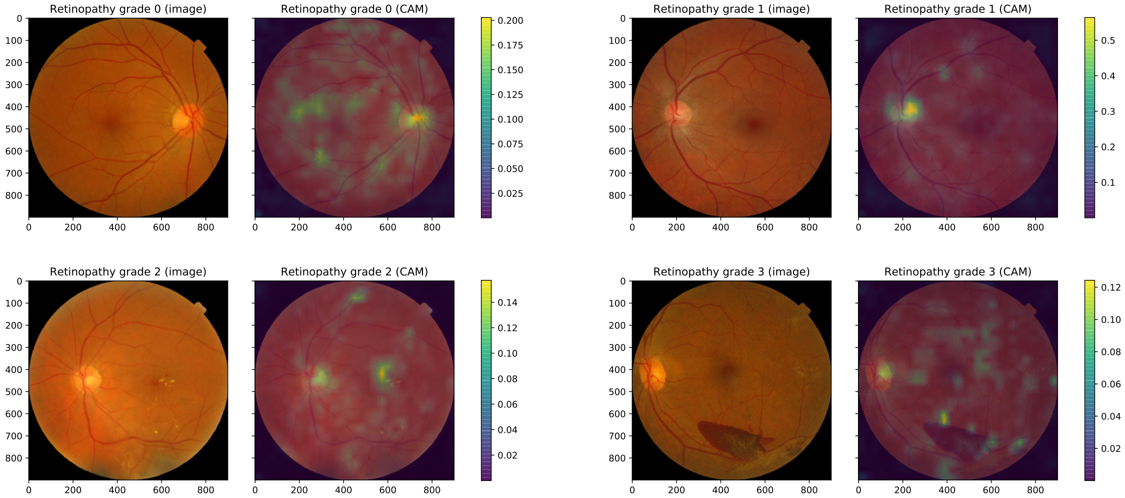

Retinopathy grade 0 (image)
Retinopathy grade 0 (CAM)
Retinopathy grade 1 (image)
Retinopathy grade 1 (CAM)
Retinopathy grade 2 (image)
Retinopathy grade 2 (CAM)
Retinopathy grade 3 (image)
Retinopathy grade 3 (CAM)
0
100
200
300
400
500
600
700
800
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.5
0.4
0.3
0.2
0.1
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.12
0.10
0.08
0.06
0.04
0.02

Figure 8: Original sample images and its corresponding Class Activation Maps (CAM) at layer 4 of the ResNet-50b model. One sample per retinopathy grade is shown.

grade 2 increases from 0.79 to 0.88, which may be indicative of the heterogeneity of grades 0 and 1 impacting the performance of this last grade.

Finally, we have also tested the different systems under a binary classification scenario, aimed to distinguish between RG=0 and RG=3. Being the most extreme cases, the performance of the system significantly increases, and even AlexNet or ZhangNet, with CR below 0.4 in the 4-class experiment, is able to distinguish between the two groups with more than 66% accuracy. Our ResNet based models achieve outstanding performance (a Correct Rate of 0.953 and AUCs over 0.95 as shown in Table 3), confirming their usefulness in this DR grading purpose. The fact that the ResNet models that retrain not only the last fully-connected layers (ResNet-50a) but also the last residual layer (ResNet-18, ResNet-50b), proves again that the finetuning of the feature extraction blocks is of fundamental importance for retinopathy grading.

In terms of interpretability, the grading decisions of the different systems can be substantiated using some graphical methodologies, such as the well-known Class Activation Maps (CAM) [44]. These maps can be used to provide a practical way to assess how different areas of the input image influence the final decision of the system, at the same time that allows a visual interpretation of the results. We display the absolute value of the CAMs, so that their intensity can roughly be interpreted as how influential are different areas of the image on the final result. The CAM for some samples belonging to all DR grades is displayed at Figure 8. While there exist a CAM for each output neuron, we display just those corresponding to maximum activation neuron when it is the right class.

Given that grade 0 should show no identifiable lesions, we could expect the higlihgted areas to be widely spread over the whole eye fundus and even outside the image. This is what happens in the first case. For grades 1-3, we observe a higher concentration of areas around typical lesions associated to retinopathy. In the case of grade 1, there is hardly any lesion highlighted, just a few darkened areas around blood vessels in the top and right part. For grades 2 and 3, there are clearer marks, especially around microaneurysms (grade 2, right part, grade 3 top part) and neovascularization areas (grade 3, right part and above the big dark area in the lower part).

Although images from Messidor database are acquired with high resolution and their quality is higher than other databases, it is specifically designed to test computer-assisted diagnosis methods of diabetic retinopathy. As a result, differences among classes are far more subtle than in other retina databases. As an example, [45] proposes the use of a modified version of GoogLeNet including regularization methods to improve the generalization capabilities, achieving a correct classification rates of 0.74, 0.68 and 0.57 for the 2-class, 3 class and 4-class classification experiments respectively. When compared to our methodology, whereas the 4-class experiment obtains similar performance, we can check that the performance of the ResNet-50b in the 3 and 2-class experiments is way higher than the reported GoogLeNet, which can be due to our transfer learning strategy, bringing together an excellent architecture with pre-trained low-level general feature extractors, proving the advantages of the proposed methodology in the automatic grading of DR.

6. Conclusions

In this work we propose an automatic Diabetic Retinopathy (DR) grading system based on residual networks. Unlike other methods in the literature trained with specific retinography datasets, we applied the transfer learning paradigm in a far more bio-inspired learning approach under the assumption that more abstract features are extracted by the first layers of a deep enough network, which could potentially increase the generalization ability of the system. The first layers of a pre-trained network have therefore the ability to extract knowledge for a wide range of applications, so all the networks shown in this paper were pre-trained using the ImageNet dataset, composed of general content images but not particularly related to retina disease diagnosis. Then, the networks were fine-tuned using retinal images from the Messidor database along with a slight data augmentation to improve the generalization capability.

Our work tests the ResNet-18 and ResNet-50b architectures, but also uses transfer learning on AlexNet, ZhangNet, SqueezeNet and two versions of VGGnet as baseline experiments under the same conditions. We tested the different networks under various multiclass and binary scenarios, including a fully 4-class grading system, a 3-class with aggregation of grades 0 and 1, and a binary classification problem using grades 0 and 3.

Our system proves its ability in differentiating among different grades of DR, with outstanding performance in detecting advanced DR (grade 3), and a grade 2 performance far above the baseline methods. It also was able to correctly identify most grade 0 images, although the system clearly improved when aggregating grades 0 and 1, mainly due to the subtlety of the lesions, as well as a cross-confusion on the labelling, increasing the AUC of grades 2 and 3 to 0.811 and 0.918. Particularly, when considering only grades 0 and 3, the classification performance increased up to 95.5% accuracy, with a sensitivity of 0.983 for grade 3. We also derived class activation maps that help to locate the lesions that helped the network in the grading procedure. The battery of tests and analyses performed on the images as well as its performance results demonstrate that the proposed system based on ResNet architectures via transfer learning could be of great use for future automatic Diabetic Retinopathy grading systems of clinical use.

Acknowledgements

This work was partly supported by the MINECO/FEDER under TEC2015-64718-R, RTI2018-098913-B-I00, PSI2015-65848-R and PGC2018-098813-B-C32 projects. We gratefully acknowledge the support of NVIDIA Corporation with the donation of one of the GPUs used for this research. Work by F.J.M.M. was supported by the MICINN "Juan de la Cierva - Formación" Fellowship.

References

[1] D. J. Pettitt, J. Talton, D. Dabelea, J. Divers, G. Imperatore, J. M. Lawrence, A. D. Liese, B. Linder, E. J. Mayer-Davis, C. Pihoker, S. H. Saydah, D. A. Standiford, R. F. Hamman, , Prevalence of diabetes in u.s. youth in 2009: The search for diabetes in youth study, Diabetes Care 37 (2) (2014) 402–408.

[2] A. A. of Ophthalmology. The Eye M.D. Association, International clinical diabetic retinppathy disease severity scale (2002). URL http://www.icoph.org/downloads/Diabetic-Retinopathy-Scale.pdf

[3] S. Charumathi, et al., Incidence and progression of diabetic retinopathy: a systematic review, The Lancet Diabetes & Endocrinology 7 (2) (2019) 140 – 149.

[4] A. Ortiz, J. M. Górriz, J. Ramírez, F. J. Martínez-Murcia, A. D. N. Initiative, et al., Automatic ROI Selection in Structural Brain MRI Using SOM 3D Projection, PLOS ONE 9 (4) (2014) e93851.

[5] A. Ortiz, J. Munilla, J. M. Górriz, J. Ramírez, Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease, International Journal of Neural Systems 26 (07) (2016) 1650025.

[6] F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, A. Ortiz, A structural parametrization of the brain using hidden markov models-based paths in alzheimer's disease, International Journal of Neural Systems 26 (07) (2016) 1650024, pMID: 27354189. arXiv:http://www.worldscientific.com/doi/pdf/10.1142/S0129065716500246,doi:10.1142/S0129065716500246. URL http://www.worldscientific.com/doi/abs/10.1142/S0129065716500246

[7] J. V. B. Soares, J. J. G. Leandro, R. M. Cesar, H. F. Jelinek, M. J. Cree, Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification, IEEE Transactions on Medical Imaging 25 (9) (2006) 1214–1222. doi:10.1109/TMI.2006.879967.

[8] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, B. van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Transactions on Medical Imaging 23 (4) (2004) 501–509. doi:10.1109/TMI.2004.825627.

[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444. doi:10.1038/nature14539.

[10] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[11] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, S. A. Barman, An ensemble classification-based approach applied to retinal blood vessel segmentation, IEEE Transactions on Biomedical Engineering 59 (9) (2012) 2538–2548. doi:10.1109/TBME.2012.2205687.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009, pp. 248–255.

[13] L. P. Cunha, E. A. Figueiredo, H. P. Araújo, L. V. F. Costa-Cunha, C. F. Costa, J. d. M. C. Neto, A. M. F. Matos, M. M. d. Oliveira, M. G. Bastos, M. L. R. Monteiro, Non-mydriatic fundus retinography in screening for diabetic retinopathy: Agreement between family physicians, general ophthalmologists, and a retinal specialist, Frontiers in Endocrinology 9 (2018) 251.

[14] N. Salamat, M. M. S. Missen, A. Rashid, Diabetic retinopathy techniques in retinal images: A review, Artificial Intelligence in Medicine 97 (2019) 168 – 188.

[15] K. Sreejini, V. Govindan, Retrieval of pathological retina images using bag of visual words and plsa model, Engineering Science and Technology, an International Journaldoi:https://doi.org/10.1016/j.jestch.2019.02.002.
URL http://www.sciencedirect.com/science/article/pii/S2215098617314994

[16] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition 29 (1) (1996) 51 – 59.

[17] D. G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2, 1999, pp. 1150–1157 vol.2. doi:10.1109/ICCV.1999.790410.

[18] T. Jabid, M. H. Kabir, O. Chae, Local directional pattern (ldp); a robust image descriptor for object recognition, in: 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010, pp. 482–487.

[19] K. Kim, H. Kim, J. Seo, A neural network model with feature selection for korean speech act classification, International journal of neural systems 14 (2005) 407–14.

[20] J. Garrido, N. R. Luque, S. Tolu, E. D'Angelo, Oscillation-driven spike-timing dependent plasticity allows multiple overlapping pattern recognition in inhibitory interneuron networks, International Journal of Neural Systems 26 (2016) 1650020. doi:10.1142/S0129065716500209.

[21] E. Gawehn, J. A. Hiss, G. Schneider, Deep learning in drug discovery., Molecular informatics 35 1 (2016) 3–14.

[22] A. Ortiz, F. Martínez-Murcia, J. Munilla, J. Gorriz, J. Ramírez, Label aided deep ranking for the automatic diagnosis of parkinsonian syndromes, Neurocomputing 330 (2018) 162–171.

[23] A. Ortiz, J. Munilla, M. Martínez-Ibañez, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, Parkinson's disease detection using isosurfaces-based features and convolutional neural networks, Frontiers in Neuroinformatics 13 (2019) 1–48.

[24] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, G. Yang, Hierarchical retinal blood vessel segmentation based on feature and ensemble learning, Neurocomputing 149 (2015) 708 – 717.

[25] F. Girard, C. Kavalec, F. Cheriet, Joint segmentation and classification of retinal arteries/veins from fundus images, Artificial Intelligence in Medicine 94. doi:10.1016/j.artmed.2019.02.004.

[26] C. Mahiba, A. Jayachandran, Severity analysis of diabetic retinopathy in retinal images using hybrid structure descriptor and modified cnns, Measurement 135 (2019) 762 – 767.

[27] C. Lam, C. Yu, L. Huang, D. Rubin, Retinal lesion detection with deep learning using image patches, Investigative Opthalmology & Visual Science 59 (2018) 590.

[28] J. Y. Choi, T. Keun Yoo, J. Gi Seo, J. Kwak, T. Taewoong Um, T. Hyungtaek Rim, Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database, PLOS ONE 12 (2017) e0187336.

[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.
URL http://arxiv.org/abs/1409.1556

[30] C. Lam, D. Yi, M. Guo, T. Lindsey, Automated detection of diabetic retinopathy using deep learning, AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2017 (2018) 147—155.

[31] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy, Procedia Computer Science 90 (2016) 200 – 205, 20th Conference on Medical Image Understanding and Analysis (MIUA 2016).

[32] C. Lian, Y. Liang, R. Kang, Y. Xiang, Deep convolutional neural networks for diabetic retinopathy classification, in: Proceedings of the 2Nd International Conference on Advances in Image Processing, ICAIP '18, ACM, New York, NY, USA, 2018, pp. 68–72. doi:10.1145/3239576.3239589.
URL http://doi.acm.org/10.1145/3239576.3239589

[33] T. Shanthi, R. Sabeenian, Modified alexnet architecture for classification of diabetic retinopathy images, Computers & Electrical Engineering 76 (2019) 56 – 64.

[34] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, Curran Associates Inc., USA, 2012, pp. 1097–1105.
URL http://dl.acm.org/citation.cfm?id=2999134.2999257

[35] J. de la Torre Gallart, A. Valls, D. Puig, A deep learning interpretable classifier for diabetic retinopathy disease grading, Neurocomputing (2019) 1–12doi:10.1016/j.neucom.2018.07.102.

[36] D. Zhang, W. Bu, X. Wu, Diabetic retinopathy classification using deeply supervised resnet, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1–6. doi:10.1109/UIC-ATC.2017.8397469.

[37] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, J.-C. Klein, Feedback on a publicly distributed database: the messidor database, Image Analysis & Stereology 33 (3) (2014) 231–234. doi:10.5566/ias.1155.
URL http://www.ias-iss.org/ojs/IAS/article/view/1155

[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[39] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European conference on computer vision, Springer, 2016, pp. 630–645.

[40] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. on Knowl. and Data Eng. 22 (10) (2010) 1345–1359.

[41] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size, arXiv preprint arXiv:1602.07360.

[42] F. J. Martinez-Murcia, J. M. Gorriz, J. Ramirez, A. Ortiz, Convolutional Neural Networks for Neuroimaging in Parkinson's Disease: Is Preprocessing Needed?, International journal of neural systems (2018) 1850035–1850035doi:10.1142/s0129065718500351.

[43] R. P. Duin, Classifiers in almost empty spaces, in: Proceedings 15th International Conference on Pattern Recognition, Vol. 2, 2000, pp. 1–7.

[44] B. Zhou, A. Khosla, L. A., A. Oliva, A. Torralba, Learning Deep Features for Discriminative Localization., CVPR.

[45] C. K. Lam, D. Yi, M. Guo, T. Lindsey, Automated detection of diabetic retinopathy using deep learning, in: AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, 2018, p. 147.