FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Diabetic Retinopathy Grading in Color Eye Fundus Images

Teresa Manuel Sá Finisterra Araújo



Submitted for the degree of Doctor of Philosophy in Electrical and Computer Engineering

Supervisor: Aurélio Campilho, Full Professor, FEUP

Co-supervisor: Ana Maria Mendonça, Associate Professor, FEUP

May 19, 2021

Diabetic Retinopathy Grading in Color Eye Fundus Images

Teresa Manuel Sá Finisterra Araújo

Submitted for the degree of Doctor of Philosophy in Electrical and Computer Engineering

May 19, 2021

Abstract

Diabetic retinopathy (DR) is a complication of diabetes and one of the leading causes of blindness worldwide. However, the majority of visual loss cases can be prevented with early detection, and thus regular check-ups are essential. DR screening programs have been developed across the world, in which detection and severity grading are performed based on the analysis of retinal lesions visible in eye fundus images. Due to the complexity of these tasks and the high workload of the specialists, computer-aided diagnosis (CAD) systems are desirable for reducing their burden and the diagnosis' subjectivity. However, the complexity of the grading task and the wide variety of lesions makes the development of these systems a non-trivial task. Deep learning approaches have shown promising performance as CAD systems, but their black-box behaviour hinders clinical application. This thesis focuses on the development of DR grading tools aiming at improving the screening pipeline.

First, we propose UOLO, a novel framework for the simultaneous detection and segmentation of structures of interest in medical images. UOLO consists of a fully-convolutional encoder-decoder segmentation module which intermediate abstract representations are processed and used as input for an one stage object detection framework. The system is optimized simultaneously for detecting a class of objects and segmenting an optionally different class of structures, using a novel loss function. We validate UOLO on the tasks of optic disc detection and segmentation and fovea detection in retinal images, which are structures of interest for the diagnosis of several eye diseases. On these tasks, the model achieves state-of-the-art performance on public datasets.

Then, DR|GRADUATE, a novel deep learning-based DR grading CAD system, is proposed. With the goal of mitigating the black-box behaviour commonly associated with deep learning, the network supports its decision by providing a medically interpretable explanation and an estimation of how uncertain that prediction is. Specifically, DR|GRADUATE was designed taking into account the ordinal nature of the DR grading problem. A novel Gaussian-sampling approach built upon a Multiple Instance Learning framework allows DR|GRADUATE to infer an image grade associated with an explanation map and a prediction uncertainty while being trained only with image-wise labels. This network was trained on the Kaggle DR detection training set and evaluated across multiple datasets. We show that high performance occurs for images with low prediction uncertainty, thus indicating that this uncertainty is a valid measure of the predictions' quality. Further, bad quality images are generally associated with higher uncertainties, showing that images not suitable for diagnosis indeed lead to less trustworthy predictions. Tests on unfamiliar medical image data types suggest that DR|GRADUATE allows outlier detection. Additionally, the attention maps generally highlight regions of interest for diagnosis. The obtained results show the great potential of DR|GRADUATE as a second-opinion system in DR severity grading.

Then, we present a novel data augmentation scheme motivated by the performance of the aforementioned model on the proliferative DR (PDR) detection task. This lower performance is associated with the small representation of this type of images as well as incongruences in the annotations, which leads to a reduced variability of the samples and thus lower generalization

capability of the model. The proposed heuristic-based data augmentation method relies on the generation of neovessel (NV)-like structures. This synthesis is based on the general knowledge of common location and shape of these structures. NVs are synthesized and introduced in pre-existent retinal images which can then be used for enlarging deep neural networks' training sets. The data augmentation scheme was tested on multiple datasets, and allows to improve the DR|GRADUATE's capacity to detect NVs.

Finally, we present a scheme for estimating the DR grade of an image based on an ensemble of the previously trained networks, and evaluate the final method on different public datasets. The method's performance is in line with recent state-of-the-art deep learning-based methods, with DR|GRADUATE being the only that inherently presents an uncertainty associated with the diagnosis and an explanation map per grade. We also perform a thorough analysis on the private SCREEN-DR dataset, both of the automatic algorithm and of the specialists involved in the annotation of the dataset. The inter-observer analysis has shown that there are cases of high disagreement, highlighting the complexity of the DR grading task and the subjectivity in diagnosis of the most difficult cases. The proposed method achieves a performance comparable with the annotators, with a particularly good performance for DR and referable DR detection, thus showing its potential as a screening tool, as it allows to discard *a priori* the healthy or mild cases and refer the most severe ones to the specialists.

Resumo

A retinopatia diabética (RD) é uma complicação da diabetes e uma das principais causas de cegueira em todo o mundo. Contudo, a maioria dos casos de perda de visão pode ser prevenida com deteção precoce, sendo essencial a verificação regular. Programas de rastreio de RD têm vindo a ser desenvolvidos em todo o mundo, nos quais a deteção e classificaçao da gravidade da doença são determinados com base na análise das lesões visíveis em imagens do fundo do olho. Devido à complexidade destas tarefas e à elevada carga de trabalho dos especialistas, sistemas de diagnóstico assistido por computador (DACo) são desejáveis para reduzir a carga de trabalho e a subjetividade no diagnóstico. Todavia, a compexidade da tarefa de estadiamento e a grande variedade de lesões fazem com que o desenvolvimento destes sistemas não seja trivial. Abordagens baseadas em aprendizagem profunda (*deep learning*) têm mostrado um desempenho promissor como sistemas DACo, mas o seu comportamento de caixa-preta dificulta a aplicação clínica. Esta tese foca-se no desenvolvimento de ferramentas de estadiamento de RD com o objetivo de melhorar o rastreio.

Primeiro, propomos a UOLO, uma nova abordagem para a deteção e segmentação simultâneas de estruturas de interesse em imagens médicas. UOLO consiste num módulo de segmentação codificador-descodificador completamente convolucional, cujas representações intermédias são processadas e usadas como entrada para uma estrutura de deteção de objetos de etapa única. O sistema é otimizado simultaneamente para detetar uma classe de objetos e segmentar uma classe de estruturas opcionalmente diferente, usando uma nova função de custo. Validamos a UOLO nas tarefas de detecção e segmentação do disco ótico e deteção da fóvea, que são estruturas de interesse no diagnóstico de diversas doenças oculares. Nestas tarefas, o modelo apresenta desempenho ao nível do estado-de-arte em conjuntos de dados (*datasets*) públicos.

De seguida, propomos DR|GRADUATE, um novo sistema DACo de classificação de DR baseado em *deep learning*. Com o objetivo de mitigar o comportamento de caixa-preta commummente associado ao *deep learning*, a rede suporta a sua decisão fornecendo uma explicação medicamente interpretável e uma estimativa de quão incerta é essa previsão. Especificamente, DR|GRADUATE foi desenhado tendo em consideração a natureza ordinal do problema de classificação de RD. Uma nova abordagem de amostragem Gaussiana construída sobre uma estrutura de aprendizagem de multiplas instâncias permite a DR|GRADUATE inferir um grau de RD associado a um mapa de explicação e a uma incerteza de previsão sendo treinado apenas com etiquetas ao nível da imagem. Esta rede foi treinada no *dataset* "Kaggle DR detection" e avaliada em vários *datasets*. Mostramos que imagens com incertezas de previsão baixas estão associadas a um desempenho elevado, o que sugere que essa incerteza é uma medida válida da qualidade das previsões. Além disso, imagens de má qualidade estão comummente associadas a incertezas mais elevadas, mostrando que imagens não adequadas para diagnóstico levam de facto a previsões menos fiáveis. Testes em tipos de dados de imagens médicas desconhecidos sugerem que DR|GRADUATE permite a detecção de amostras fora da distribuição. Além disso, os mapas de atenção destacam geralmente regiões de interesse para diagnóstico. Os resultados obtidos mostram o grande poten-

cial de DR|GRADUATE como sistema de segunda opinião para a classificação da gravidade de RD.

Depois, apresentamos um novo esquema de aumento de dados motivado pelo desempenho do modelo acima mencionado na tarefa de deteção de RD proliferativa (RDP). Este desempenho inferior está associado à baixa representação deste tipo de imagens e também a incongruências nas anotações, o que leva a uma menor variabilidade das amostras e, portanto, a uma menor capacidade de generalização por parte do modelo. O método de aumento de dados proposto baseia-se numa abordagem heurística para a geração de estruturas semelhantes a neovasos (NV). Esta síntese é baseada no conhecimento geral da localização e forma comuns destas estruturas. Os NVs são sintetizados e introduzidos em imagens da retina pré-existentes, que podem então ser usadas para ampliar os *datasets* de treino de modelos *deep learning*. O esquema de aumento de dados foi testado em vários *datasets* e permite melhorar a capacidade de DR|GRADUATE para detetar NVs.

Finalmente, apresentamos um esquema para estimar o grau de RD de uma imagem com base na combinação das previsões das redes previamente treinadas e avaliamos o método final em diferentes *datasets* públicos. O desempenho do método está em linha com o estado-de-arte recente baseado em *deep learning*, sendo DR|GRADUATE o único que apresenta inerentemente uma incerteza associada ao diagnóstico e um mapa de explicação por grau. Também realizamos uma análise minuciosa no *dataset* privado SCREEN-DR, tanto do algoritmo automático como dos especialistas envolvidos na anotação. A análise inter-observador mostrou que há casos de discordância elevada, evidenciando a complexidade da tarefa de estadiamento de RD e a subjetividade no diagnóstico dos casos mais difíceis. O método proposto apresenta um desempenho comparável aos anotadores, com desempenho particularmente bom para deteção de RD e RD referenciável, mostrando assim o seu potencial como ferramenta de triagem, uma vez que permite descartar *a priori* os casos saudáveis ou ligeiros e encaminhar os mais graves para o especialistas.

Acknowledgments

Teresa Araújo is funded by the FCT grant contract SFRH/BD/122365/2016. This work is also financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia within project CMUP-ERI/TIC/0028/2014. The author would like to thank the collaboration of the C-BER Biomedical Imaging Lab from INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Departamento de Oftalmologia do Centro Hospitalar Universitário São João, Departamento de Oftalmologia do Hospital de Braga, Faculdade de Medicina da Universidade do Porto, ARSN - Administração Regional de Saúde do Norte, i3S - Instituto de Investigação e Inovação em Saúde, and the Diagnostic Image Analysis Group from RadboudUMC.



Contents

1 Introduction **1**
- 1.1 Anatomical context . . . 1
  - 1.1.1 Eye and retina . . . 1
  - 1.1.2 Eye fundus images . . . 1
- 1.2 Diabetic retinopathy . . . 2
- 1.3 Diabetic retinopathy screening . . . 4
- 1.4 Thesis Objectives . . . 4
- 1.5 Contributions . . . 5
- 1.6 Thesis outline . . . 6

2 State-of-the-art review **7**
- 2.1 Scientific impact of diabetic retinopathy automatic diagnosis . . . 7
- 2.2 Methodologies' fundamentals . . . 8
  - 2.2.1 Convolutional neural networks . . . 8
  - 2.2.2 Uncertainty of the prediction . . . 12
  - 2.2.3 Explainability of the decision . . . 14
- 2.3 Methods for DR grading . . . 16
  - 2.3.1 Lesion-based methods . . . 17
  - 2.3.2 Non-lesion-based methods . . . 21
  - 2.3.3 Relevant research contributions . . . 22
  - 2.3.4 Comparative overview of the existent techniques . . . 26
- 2.4 Retinal imaging . . . 27
  - 2.4.1 Datasets for DR grading . . . 27
- 2.5 Concluding remarks . . . 29

3 UOLO - automatic object detection and segmentation in biomedical images **31**
- 3.1 Introduction . . . 31
  - 3.1.1 Optic disc and fovea detection and segmentation . . . 32
- 3.2 UOLO framework . . . 34
  - 3.2.1 Object segmentation module . . . 34
  - 3.2.2 Object detection module . . . 35
  - 3.2.3 UOLO for joint object detection and segmentation . . . 36
- 3.3 Experiments and results . . . 37
  - 3.3.1 Datasets and experimental details . . . 37
  - 3.3.2 Results and discussion . . . 38
- 3.4 Concluding remarks . . . 41

4 DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading 43
4.1 Introduction . . . 44
4.1.1 State-of-the-art . . . 45
4.1.2 Contributions . . . 47
4.2 DR|GRADUATE for diabetic retinopathy (DR) grading . . . 47
4.2.1 Predicting a grade with uncertainty . . . 48
4.2.2 Explainability of the DR grade . . . 50
4.2.3 Dealing with class imbalance . . . 51
4.3 Experiments . . . 52
4.3.1 Datasets . . . 52
4.3.2 Evaluation . . . 54
4.3.3 Parameter setting . . . 56
4.3.4 Training details . . . 57
4.4 Results . . . 57
4.4.1 DR grading . . . 57
4.4.2 Uncertainty estimation . . . 58
4.4.3 Explainability . . . 61
4.5 Discussion . . . 63
4.5.1 DR grading . . . 63
4.5.2 Uncertainty estimation . . . 66
4.5.3 Explainability . . . 69
4.6 Concluding remarks . . . 70

5 Data augmentation for improving proliferative diabetic retinopathy detection 71
5.1 Introduction . . . 72
5.2 Methods . . . 73
5.2.1 Insertion location . . . 74
5.2.2 Quantity . . . 75
5.2.3 Semi-random neovessel generation . . . 75
5.2.4 Color assignment and neovessel insertion . . . 79
5.2.5 DR Grading . . . 80
5.3 Results . . . 81
5.3.1 Datasets . . . 81
5.3.2 Results and discussion . . . 81
5.4 Concluding remarks . . . 85

6 Analysis of the performance of specialists and an automatic algorithm in DR grading 87
6.1 DR grading prediction . . . 87
6.2 Performance evaluation . . . 88
6.3 SCREEN-DR analysis . . . 90
6.3.1 Available annotation data . . . 91
6.3.2 Intra-observer analysis . . . 91
6.3.3 Inter-observer analysis . . . 91
6.3.4 Automatic method and observers' analysis . . . 97
6.4 A note on DR grading . . . 104
6.5 Concluding remarks . . . 104

7 Conclusions 107

8 List of publications **111**

A Methods for diabetic retinopathy diagnosis **115**

References **119**

List of Figures

1.1 Anatomy of the eye. 2
1.2 Examples of color fundus images with different grades of diabetic retinopathy (DR) 3
1.3 Diabetic retinopathy (DR) screening pipeline incorporating automatic methods for DR assessment. 5

2.1 Number of published scientific papers on automated diagnosis of diabetic retinopathy from 2005 to 2020. 8
2.2 Examples of uncertainty estimation schemes. 13
2.3 Examples of explanation map generation techniques. 15
2.4 General approaches for diabetic retinopathy grading 17

3.1 Using UOLO for fovea detection and optic disc detection and segmentation. . . . 34
3.2 UOLO framework, nesting an U-Net responsible for segmentation and feature extraction for an YOLOv2-based detector. 35
3.3 Examples of results of UOLO on Messidor images. 38
3.4 Examples of UOLO results. 39

4.1 Examples of typical retinal lesions characteristic of diabetic retinopathy. 44
4.2 DR|GRADUATE's architecture. 48
4.3 Uncertainty estimation and loss representation. 50
4.4 Attention maps construction. 51
4.5 Class distribution of the DR grading datasets used for evaluation. 52
4.6 Examples of ophthalmologist's annotations on the SCREEN-DR dataset. 53
4.7 Confusion matrices of DR|GRADUATE's predictions *vs* the ground truth DR grades for different test datasets. 57
4.8 Uncertainty of the predictions in relation with the model's performance. 58
4.9 Uncertainty-associated predictions for images from the Kaggle DR detection test set. 59
4.10 Uncertainty prediction for different types of tasks. 60
4.11 Examples of convolution with Gaussian filters of increasing standard deviation. . 60
4.12 Uncertainty estimation of DR|GRADUATE for images resulting from convolution with Gaussian filters of increasing standard deviation. 61
4.13 Maps predicted by our method along with the ophthalmologist's annotated lesions in the SCREEN-DR dataset. 62
4.14 O_{max} and O_{class} for different threshold values applied to the attention maps 62
4.15 Examples of images from Messidor-2 and Kaggle DR detection test sets that are predicted by DR|GRADUATE as R4. 63
4.16 Examples of R0-R2 Kaggle DR detection dataset images (not used for training) along with the DR|GRADUATE predictions. 65

4.17 Examples of R2 and R3 Kaggle DR detection dataset images (not used for training) along with DR|GRADUATE predictions. 65
4.18 Examples of healthy Messidor-2 images mispredicted as R1. 66
4.19 Example of a pair of good/bad quality images from the HRF dataset. 68

5.1 Pipeline of the proposed neovessel generation algorithm and insertion in the retinal image. 74
5.2 Semi-random neovessel generation algorithm's parameters example. 76
5.3 Examples of step-wise generation of artificial neovessels for the different considered shapes. 78
5.4 Color attribution to a generated neovessel (NV). 80
5.5 Examples of images with synthesized artificial neovessels (NVs). 82
5.6 Confusion matrices of the new and initial (DR|GRADUATE's) predictions *vs* the ground truth DR grades in the SCREEN-DR, Kaggle DR detection test and Messidor-2 datasets. 83
5.7 Maps predicted by our method and DR|GRADUATE along with the ophthalmologist's annotated lesions in the SCREEN-DR dataset. 84
5.8 Maps predicted by our method along with the ophthalmologist's annotated lesions in the SCREEN-DR dataset in cases in which the model mispredicts R4-labeled images. 85

6.1 Confusion matrices of DR|GRADUATE's predictions *vs* the ground truth DR grades for different test datasets. 88
6.2 Confusion matrices of published DR grading methods in the Kaggle test set. . . . 90
6.3 Annotation times (in seconds) vs DR grades. 92
6.4 Percentage of partial quality images (out of the partial and good quality images) accordingly to the attributed DR grade. 92
6.5 Agreement between annotators. Confidence stands for the number of annotators that agree with the most common label (majority voting). 93
6.6 Examples of images with complete agreement between all four annotators. 94
6.7 Examples of images with high disagreement between annotators. 95
6.8 Annotation times *vs* annotation disagreement. 96
6.9 DR grading performance of each annotator comparing to each one of the other three. 97
6.10 DR grading performance of each annotator comparing to each one of the other three. 97
6.11 Uncertainty of the algorithm's prediction *vs* annotated image quality. 98
6.12 Outliers of uncertainty *vs* image quality plot distributed accordingly to the DR grade attributed by each annotator. 98
6.13 Examples of outlier images from the uncertainty *vs* image quality plot for the annotator *538a*. 99
6.14 Uncertainty of the algorithm's prediction *vs* the annotator's disagreement in the diagnosis for each pair of annotators. 100
6.15 DR grading performance of the automatic method having as reference each of the annotators. 100
6.16 Quadratic weighted kappa *vs* uncertainty (*xx*-axis in logarithmic scale) 101
6.17 DR grading performance of the automatic method (quadratic weighted kappa score) having as reference each of the annotators accordingly to the image quality. . . . 101

6.18 DR grading performance of the annotator (quadratic weighted kappa score), having as reference each one of the other annotators, accordingly to the image quality label. 101
6.19 Per class performance of the algorithm and the annotators having as reference each of the annotators. 102
6.20 DR and referable DR performance of the algorithm and the annotators in terms of precision, recall, specificity and F1-score. 103

List of Tables

2.1 Diabetic retinopathy severity scale used in Akram *et al.* (2014). 19
2.2 Diabetic retinopathy severity scale from Messidor dataset annotation. 19
2.3 International clinical Diabetic retinopathy (DR) disease severity scale. 20
2.4 Modified Davis grading used in Takahashi *et al.* (2017) for the DMR dataset annotation. 22
2.5 Publicly available datasets of color eye fundus images 29

3.1 UOLO performance on optic disc detection and segmentation and fovea detection. 38
3.2 State-of-the-art for OD detection and segmentation and FV detection. 40

4.1 Diabetic retinopathy severity scale. 45
4.2 Quadratic weighted kappa for DR|GRADUATE's DR grading in four different datasets. 57
4.3 Assessment of the produced attention maps (threshold=0.3) in comparison with specialist annotated lesions. 61

5.1 Ranges and values of the semi-random generation algorithm parameters for each of the considered neovessel shapes. 77
5.2 Quadratic weighted kappa for the new and initial (DR|GRADUATE's) DR grading in the SCREEN-DR, Kaggle DR detection test and Messidor-2 datasets. 81

6.1 Quadratic weighted kappa for DR|GRADUATE's DR grading in four different datasets. 89
6.2 Performance of published methods for DR grading. 90
6.3 DR grading performance of the annotators and the automatic method. 100
6.4 Area under the curve (AUCs) for each detection task (ROCs from Fig. 6.19). . . 103

A.1 Methods for diabetic retinopathy grading. 115

Abbreviations and Symbols

Abbreviations

AUC area under the curve.

CAD computer-aided diagnosis.

CNN convolutional neural network.

CWS cotton wool spot.

DR diabetic retinopathy.

ETDRS Early Treatment Diabetic Retinopathy Study.

EX hard exudate.

FAZ foveal avascular zone.

FOV field of view.

FV fovea.

GLCM gray level co-occurence matrix.

GMM Gaussian mixture model.

HEM hemorrhage.

IoU intersection over union.

IRMA intraretinal microvascular abnormality.

kNN k-nearest neighbors.

LBP local binary patterns.

MA microaneurysm.

ME macular edema.

MIL multiple instance learning.

NN neural network.

NPDR non-proliferative diabetic retinopathy.

NV neovascularization.

OD optic disc.

PDR proliferative diabetic retinopathy.

PNN probabilistic neural network.

ROI region-of-interest.

SVM support vector machine.

UWF Ultrawide-field.

Chapter 1

Introduction

Diabetic retinopathy (DR) is a complication of diabetes and one of the leading causes of blindness worldwide. The majority of visual loss cases can be prevented with early detection and adequate treatment. Consequently, regular check-ups are essential to achieve an early diagnosis. DR screening programs have been developed across the world, in which DR detection and DR severity grading are performed, based on the analysis of retinal lesions visible in color eye fundus images. Due to the complexity of this task and high workload of the specialists, computer-aided diagnosis systems are desirable for reducing their burden and the diagnosis' subjectivity. This thesis focuses on the development of DR grading tools aiming at improving the DR screening pipeline.

In this chapter, the anatomical context of the problem is described, as well as the diabetic retinopathy prevalence around the world, the disease stages and progression, and the DR screening pipeline. Then, the main objectives and scientific contributions of this thesis are detailed, followed by the presentation of the document organization.

1.1 Anatomical context

1.1.1 Eye and retina

The retina is a thin transparent tissue that constitutes the third and most internal layer of the eye. It is located between the vitreous (interiorly) and the choroid layer (externally) (Fig. 1.1). The retina is a light-sensitive tissue that converts the incoming light into neural signals that can be understood by the brain. With the adequate techniques, such as fundus photography, the retina is visible from the outside, making it accessible for non-invasive imaging.

1.1.2 Eye fundus images

Fundus photography is a well established technique for retinal imaging, widely used for screening and detection of several eye and systemic diseases. This technique consists in capturing a photograph of the back of eye (*i.e.*, fundus). Specifically, in these images the intensity represents the amount of reflected light of a specific waveband (Abràmoff *et al.* (2010)). Fundus cameras, used

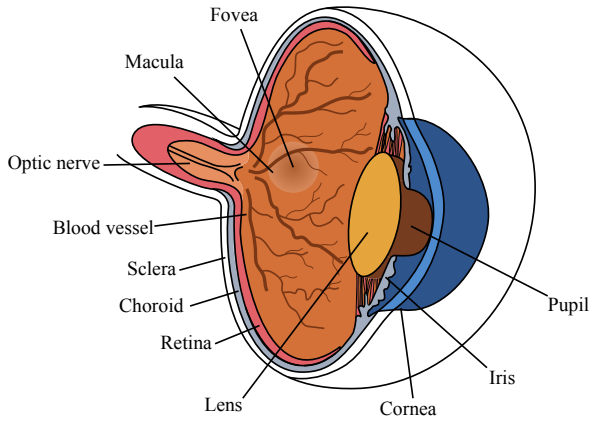


Figure 1.1: Anatomy of the eye.

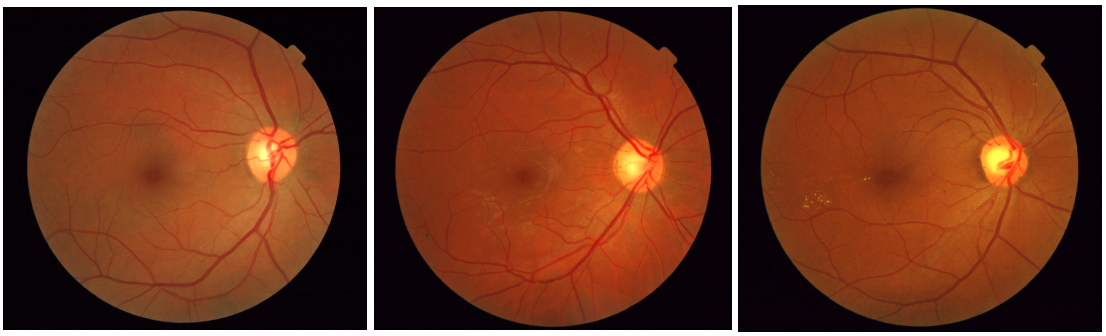
for capturing these images, are characterized by the angle of view, *i.e.*, the field of view (FOV), which standard value is 30°. However, wide angle fundus cameras capture images between 40° and 140°, providing proportionately less retinal magnification (Hipwell (1997)).

Eye fundus images allow to distinguish the following main structures: 1) fovea (FV), a reddish spot with no vessels located at the center of the macula; 2) optic disc (OD), which is a round structure in the center of the retina that corresponds to the entry point of the optic nerve and vessels that supply the retina, and 3) blood vessels, that radiate from the OD center.

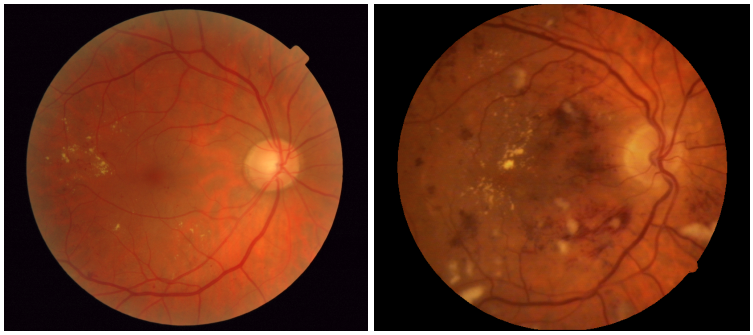
1.2 Diabetic retinopathy

DR is a complication of diabetes and is one of the leading causes of blindness worldwide (The Royal College of Ophthalmologists, 2012; Wild, 2004). According to Schully (2012), the number of diabetic patients in 2012 was of 346 million people. This number is expected to increase to over 552 million people in 2030 (The Royal College of Ophthalmologists, 2012). In 2015, the prevalence of diabetes in the Portuguese population with ages between 20 and 79 years was of 1 million, which corresponds to 13.3% of the people in this age group (Sociedade Portuguesa de Diabetologia, 2016). The majority of visual loss cases can be prevented with early detection and adequate treatment (Garg and Davis, 2009; Wu *et al.*, 2013). However, one third of the diabetic patients suffer from DR without having symptoms of visual problems, leading to the progression of the disease without treatment (Acharya *et al.*, 2012). Consequently, regular check-ups are essential to achieve early diagnosis of DR. DR screening programs have been developed across the world, for DR detection and DR severity grading, based on the analysis of retinal lesions visible in color eye fundus images.

DR is generally classified according to the absence or presence of new abnormal vessels as non-proliferative diabetic retinopathy (NPDR) or proliferative diabetic retinopathy (PDR), respectively. NPDR can be further graded as mild, moderate and severe (The Royal College of Ophthalmologists, 2012). Referable DR is considered to be moderate NPDR or worse, meaning that



(a) No signs of DR. **(b)** Mild NPDR. **(c)** Moderate NPDR.



(d) Severe NPDR. **(e)** PDR.

Figure 1.2: Examples of color fundus images with different grades of diabetic retinopathy (DR). NPDR: non proliferative DR; PDR: proliferative DR.

healthy and mild NPDR are non referable cases. Fig. 1.2 shows examples of images from each of these DR severity degrees/stages. Mild NPDR is characterized by the presence of microaneurysms (MAs), which are the first signs of DR, and appear as reddish small and circular dots (Zhang *et al.*, 2010), consisting in focal dilations of the retinal capillaries (Mahendran and Dhanasekaran, 2015). Hemorrhages (HEMs) are bright red spots of variable appearance (Kar and Maity, 2018) which start to appear with the evolution of the disease in the deeper layers of the retina, and are caused by retinal ischemia and rupture of fragile blood vessels. The disease progression results in an increase of MAs, dot-hemorrhages (red lesions), retinal thickening, hard exudates (EXs), and the appearance of intraretinal microvascular abnormalities (IRMAs), vein beading and cotton wool spots (CWSs) (soft exudates) in the innermost retina. Exudates are intraretinal fluid deposits that contain lipids, proteins and cellular debris which result from the leakage of fluids from the blood vessels. This lesion type can be classified as hard exudates (EXs), which appear in the preliminary stage and are more yellowish, bright, and have well defined borders, and soft exudates (CWSs), which appear in a more severe stage (Mahendran and Dhanasekaran, 2015), and are more pale and without defined borders (Prentasic and Loncaric, 2016). Finally, PDR results from abnormal vessel formation, *i.e.*, neovascularizations (NVs). These vessels are thin and weak, exhibiting irregular appearance, and they do not bifurcate like normal vessels (Gupta *et al.*, 2017).

1.3 Diabetic retinopathy screening

DR screening programs have been developed across the world, namely in the United Kingdom*, Ireland†, Netherlands‡ (Abràmoff and Suttorp-Schulten, 2005), France§ (Massin *et al.*, 2008), United States¶ (Cuadros and Bresnick, 2009), Canada (Nguyen and Wong, 2009) and Portugal (Dutra Medeiros *et al.*, 2015). These programs are essential to identify cases of the disease in a timely manner and prevent its progression.

In Portugal, the Regional Health Administration (ARS) are the entities responsible for managing and implementing the DR screening process in the different regions. The DR screening is recommended to diabetic patients, who should perform exams yearly. The acquired retinographies are stored in the ARS archive, being the first medical decision taken at the reading center, where ophthalmologists analyze the retinographies and conclude about their normality. If they are normal, a re-screening of the patient is recommended after one year. If they are abnormal, the ophthalmologist determines the severity of the condition and the patient is referred for treatment (Direção Geral da Saúde, 2018). It may also be the case that the acquired images do not present enough quality for diagnosis, and the acquisition must be repeated, with the patient having to return to the health center. Current screening practice implies that ophthalmologists at the reading center analyze all images, including those of poor quality, impossible to diagnose, and images without signs of pathology, which represent about 80% of the total number of acquired images.

Given the high prevalence of DR and the limited number of specialists, automated methods are a feasible and desirable option for DR screening, reducing the burden (Scotland *et al.*, 2007) and the subjectivity in the diagnosis, and providing a second opinion to specialists. In fact, the initial phases of the screening, including the normality detection, could be fully automated. Fig. 1.3 shows an example of a pipeline which incorporates computer-aided diagnosis (CAD) systems in the different image analysis tasks from the screening process, namely in the image quality assessment, referable detection and DR grading.

The vision of the SCREEN-DR project, in which this thesis is framed, is to create a distributed and automatic mass screening platform for DR, enabling immediate response of health carers, allowing accurate follow-up strategies, and fostering technological innovation.

1.4 Thesis Objectives

As discussed, the development of automatic diagnosis systems is of great interest in DR screening programs since it reduces both specialists' burden and diagnosis' subjectivity. These systems should include a DR grading module to determine the disease stage and the adequate treatment.

*https://www.gov.uk/topic/population-screening-programmes/diabetic-eye
†https://www.diabeticretinascreen.ie
‡http://www.eyecheckklant.nl
§reseau-ophdiat.aphp.fr
¶https://www.eyepacs.com

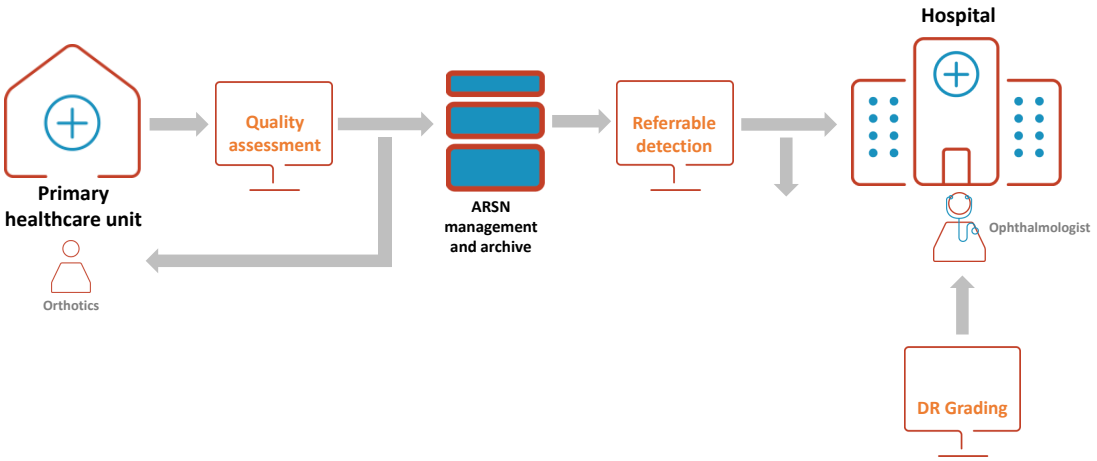


Figure 1.3: Diabetic retinopathy (DR) screening pipeline incorporating automatic methods for DR assessment.

However, the complexity of the grading task and the wide variety of lesions' characteristics makes the development of these systems a non-trivial task.

The state-of-the-art automatic DR grading methods generally have limitations, with the majority not being suitable for real-world application. Despite the promising results, most of these methods either presents low generalization capability, is not extensively and thoroughly validated, does not provide an explanation for its decisions, or produces deterministic results.

The goal of this thesis is to develop methodologies for a system that can be more easily integrated in real-world practice, that produces outcomes that can be more easily understood and possibly accepted by the health professionals. For this, its decisions should be non deterministic, associating a confidence to each diagnosis, and also a visual explanation of that decision. Additionally, its performance should be validated across multiple datasets and types of data to ensure good generalization capability.

1.5 Contributions

The main contributions of this PhD thesis are the following:

- UOLO, a deep learning-based method for automatic object detection and segmentation in biomedical images, validated for the tasks of optic disc and fovea detection and optic disc segmentation in eye fundus images, which achieves state-of-the-art performance (Araújo *et al.*, 2018a);

- DR|GRADUATE, an uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images, which provides specialists an explanation and an uncertainty associated with the diagnosis, allowing them to measure how much the algorithm's decision should be trusted (Araújo *et al.*, 2020c);

- a data augmentation scheme for improving the detection of proliferative diabetic retinopathy in retinal images, aiming at mitigating the lack of the severest grade images in the datasets for training deep learning models (Araújo *et al.*, 2020b);
- validation of the proposed DR grading framework in multiple public and in the private SCREEN-DR datasets, and assessment of the performance of the automatic method in comparison with several ophthalmologists.

1.6 Thesis outline

This document is organized as follows: Chapter 2 provides an overview of the state-of-the-art methods and datasets for diabetic retinopathy diagnosis, as well as theory fundamentals most relevant for the solutions proposed in this thesis. In Chapter 3, a method for object detection and segmentation in biomedical images is presented, and validated on the task of detecting and segmenting the optic disc and fovea in eye fundus images. Chapter 4 describes a novel methodology for DR grading (DR|GRADUATE), and assesses its performance not also for the grading task, but also in terms of validity of its uncertainty estimation and visual explanation. Then, Chapter 5 presents a data augmentation based on artificial generation of proliferative DR typical lesions, aiming at improving the performance of DR|GRADUATE in detecting this DR grade. Chapter 6 provides an overall validation of the proposed method in public and private datasets, as well as an inter-observer analysis in the private SCREEN-DR dataset. Finally, a general discussion and conclusions are presented in Chapter 7.

Chapter 2

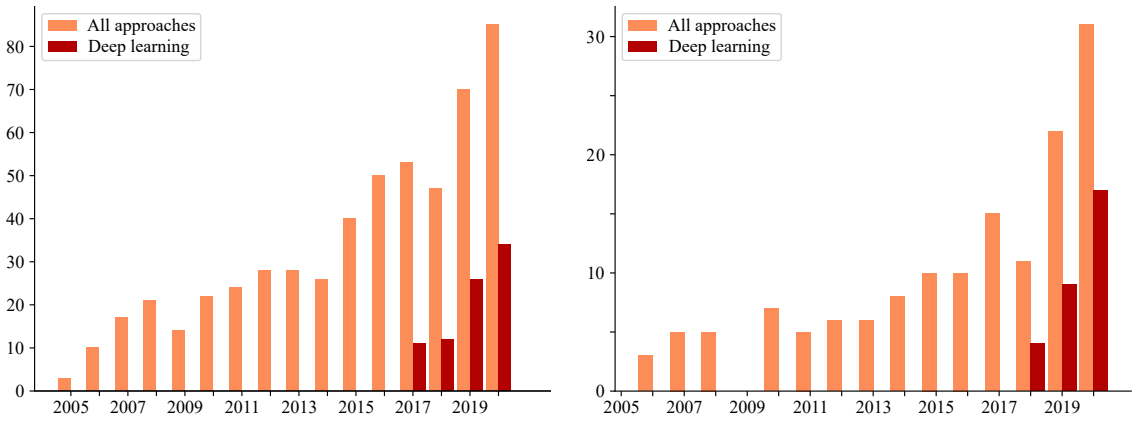
State-of-the-art review

This chapter reviews the most commonly used techniques for the development of systems for automatic DR diagnosis, focusing on grading. First, Section 2.1 shows the scientific impact of automatic DR diagnosis methods, as well as the increasing presence of deep learning approaches on these fields. Based on this, Section 2.2 introduces the fundamentals of deep learning relevant to understand the contents of this thesis and of the reviewed studies, and builds on the importance of techniques for mitigating its black-box behaviour. Then, Section 2.3 summarizes relevant contributions for DR detection and grading. Generically, the approaches were divided accordingly to their classification pipeline and the number and type of grades considered. In the end of the section a more detailed description of the most important works for this thesis is presented. An overview of the existing datasets relevant for DR grading is provided in Section 2.4.1. Finally, Section 2.5 concludes this chapter.

2.1 Scientific impact of diabetic retinopathy automatic diagnosis

The impact of diabetic retinopathy in modern society has motivated the scientific community to develop automatic algorithms for diagnosing DR on retinal images. Fig. 2.1 shows the number of published papers on the topic between 2005 and 2020, accordingly to a search performed on PubMeD (National Center for Biotechnology Information, 2020). First, the abstracts and titles of all papers that matched the search terms *diabetic retinopathy* in the referred years were retrieved. Only the hits that contained the words *CAD*, *computer-aided*, *automatic*, *algorithm* or *artificial intelligence* were kept. This list was further filtered to check for the presence of *fundus/retinal image/imaging/photography*. DR grading papers were then selected according to the presence of the words *grading*, *severity* or *grade*. Deep learning papers were counted by assessing the presence of the expressions: *deep learning*, *convolutional neural network* or *CNN*.

Accordingly to Fig. 2.1, deep learning approaches on the topic of automatic DR diagnosis have increased over the past four years. In fact, in 2020 deep learning has been used on at least one third of the automated DR diagnosis papers, and in more than 50% in DR grading papers. This increasing tendency of application of deep learning approaches, namely convolutional neural



(a) DR diagnosis. **(b)** DR grading.

Figure 2.1: Number of published scientific papers on automated diagnosis of diabetic retinopathy from 2005 to 2020. Information extracted from PubMed (National Center for Biotechnology Information, 2020).

networks (CNNs) (LeCun *et al.*, 1998), is closely related with the ease to set up, the increase of computational power over the past years, and the low requirement of applied field of knowledge, specially when compared with handcrafted feature approaches (Aresta *et al.*, 2019c). Besides, the same CNNs architectures can be applied on different tasks and imaging modalities, being a highly versatile solution to create computer-aided diagnosis' systems. Additionally, CNNs allow to use generic features learned for natural image applications (Russakovsky *et al.*, 2015) as a starting point for training, reducing the amount of medical data required to train these systems, which is usually hard to obtain. Thanks to this, CNNs are now the state-of-the-art approach for several computer vision problems and show high promise in the field of medical image analysis (Litjens *et al.*, 2017; Krizhevsky *et al.*, 2012; Abràmoff *et al.*, 2016; Gulshan *et al.*, 2016; Aresta *et al.*, 2019c).

Following the highlighted trends, the studies presented in this thesis also use as backbone deep learning approaches. With this in mind, the next section will briefly discuss some fundamentals of convolutional neural networks, including their architecture, training, and current shortcomes.

2.2 Methodologies' fundamentals

2.2.1 Convolutional neural networks

CNNs (LeCun *et al.*, 1998) are a specific type of neural networks (NNs), capable of learning feature representations to solve a given classification problem. A common example of NNs are the feedforward neural networks (FNNs), which establish a mapping function by learning the parameters so that an output can be predicted from an input. This mapping is typically the result of several neurons/perceptrons organized in layers. These neurons are interdependent units, and consist in binary classifiers that perform the mapping via a non-linear transformation of linearly combined weighted features (the activation). These nets have three fundamental layer types: 1) input layer (features to be processed); 2) hidden layers (intermediary layers of neurons, being that each layer

learns a feature transformation from the previous layer); 3) output layer (usually composed of as many neurons as the number of classes of the problem).

CNNs, despite sharing the same overall structure of FNNs, account for the spatial relation of the input. The matrix multiplication in the hidden layers of FNNs is now replaced by convolutions. These convolutions allow to learn the filters responsible for extracting features from the input image. The main types of hidden layers that constitute CNNs are: 1) convolutional layers; 2) pooling layers; 3) fully-connected layers. Convolutional layers are the main building block of the CNN. Each convolutional layer is responsible for learning kernels that when convolved with the layer's input produce new features. They are composed of neurons that perform convolutions of a given region of the previous layer. These layers have 3 dimensions: width, height and depth. The depth controls the number of kernels to be learned by that layer. Neurons are distributed over the width and height of the layer. The area analyzed by the neuron is referred as receptive field. The sets of features outputed by a convolutional layer are usually called activations. By their turn, pooling layers allow to reduce the complexity of the problem since they select a subset of activations from the previous layer. A commonly used pooling operation is the max-pooling, with takes the maximum of the activations, but the average or mininum are also used. Pooling increases the receptive field of the subsequent convolutional layers, thus allowing multi-scale analysis of the image. Blocks of convolutional-pooling layers commonly compose the main feature extraction part of CNNs. Then, usually, fully-connected layers process the previously extracted features to produce a prediction.

Network optimization/training is performed via backpropagation of the gradient of the prediction error, computed using a suitable loss function, which should be differentiable and adequate to the problem at hands. A very common loss function for classification tasks is the cross-entropy loss (Eq. 2.1), which aims at increasing the model's output probability of the correct class. When dealing with regression problems, usually the mean squared error (MSE) (Eq. 2.2) or a variation of this error is used, which penalizes predictions accordingly to the distance to the ground truth value:

$$\text{Loss}_{CE} = -\sum_{c=1}^{M} y_{i,c} log(p_{i,c}) \qquad (2.1) \qquad\qquad \text{Loss}_{MSE} = (y_i - \hat{y}_i)^2 \qquad (2.2)$$

where $y_{i,c}$ is a binary variable which indicates whether the class c is the correct class for sample i, $p_{i,c}$ is the output probability for class c, y_i and $\hat{y}_i$ are the continuous ground truth and prediction, respectively.

CNN performance is highly dependent on the architecture of the network, as well as on the optimization of hyperparemeters such as the learning rate. Training CNNs is a complex task, requiring large amounts of labeled data to avoid overfitting (Tajbakhsh *et al.*, 2016). This is particularly costly and difficult to obtain in the medical field, where expert annotation is expensive and the diseases are often scarcely represented (Tajbakhsh *et al.*, 2016). Weight regularization and dropout can be used for avoiding this, as well as data augmentation. Weight regularization is usually done by penalizing large weights on the loss function to avoid excessive adjustment to the samples of the training data. Dropout randomly deactivates a percentage of neurons during

the training (Srivastava *et al.*, 2014). Data augmentation largely increases the training samples by applying rotations, translations, shears, among others, to the input image, thus increasing the robustness of the model.

CNNs are able to recognize and make associations between complex visual pattern, similarly to the human visual system. Features learned to solve a given problem can thus be useful to another one, a process referred as transfer learning (Yosinski *et al.*, 2014). In transfer learning, a base network is trained on a base dataset and task, and then the features are transferred to a second network to be trained on the target dataset and task. This works if the features are general enough, instead of being specific to the base task. Studies suggested that higher layers of CNN computed features that are fairly general (Zeiler and Fergus, 2014; Yosinski *et al.*, 2014). The usual approach is to train a base network and copy its first n layers to the first n layers of the target network, and the remaining layers are initialized randomly. The copied layers can be frozen, *i.e.*, they do not change during training, or finetuned, *i.e.*, the errors from the target task are backpropagated through the base features (Yosinski *et al.*, 2014).

Finetuning is thus a common strategy in medical image analysis to overcome the lack of large training sets, in order to avoid overfitting to the training data (Tajbakhsh *et al.*, 2016), commonly using the base task and dataset from the ImageNet natural image analysis challenge (Russakovsky *et al.*, 2015). For instance, this was the case in the BACH challenge (Aresta *et al.*, 2019c), organized during the development of this thesis, which aimed the development of automatic breast cancer classification in histology images. Participants opted mainly for pre-trained networks that have historically achieved high performance on the ImageNet dataset (Russakovsky *et al.*, 2015), namely the VGG (Simonyan *et al.*, 2014), Inception (Szegedy *et al.*, 2015), ResNet (He *et al.*, 2016) and DenseNet (Huang *et al.*, 2017). VGG is composed of blocks of 2–3 convolutional layers with a large number of filters (3×3 kernels), followed by a max pooling layer. Blocks with the same filter size are applied multiple times in order to extract more complex and representative features. The final prediction is produced by a set of fully-connected layers. Given its linear structure and large number of parameters (around 140 million for 16 layers), it is not possible to significantly increase the depth of the model, since it would lead to huge computational requirements, and there is a tendency for overfitting (Aresta *et al.*, 2019c).

The Inception network, building on the idea that most activations in a CNN are unnecessary (*i.e.*, are zero or redundant), and thus a more efficient architecture will have sparse connections between the activations. The number of parameters (23 million parameters for 159 layers) is thus reduced by using locally sparse building blocks, *i.e.*, inception blocks. At each block the number of feature maps from the previous one is reduced via a 1×1 convolution. These features are convolved in parallel with kernels of increasing size, allowing to capture details at different scales. Additionally, the fully-connected layers at the end are replaced by global average pooling which also contributes to the reduction of the number of parameters.

Despite outperforming other networks in many applications, deep architectures are difficult to train due to the high number of parameters that need to be learned. Indeed, it has been noticed that networks with more layers, which are expected to have at least the same capacity as shallower

models, end up stabilizing their training error on higher values (He *et al.*, 2016). This phenomena is commonly referred to as training degradation. A possible strategy to solve this is the use of residual learning. In ResNet, layer inputs are shortcut to the following layers, in order to guarantee that all layers have enough information for training. In each residual block the input activation map is summed to the output of a set of convolutional layers, thus easing the flow of information. A ResNet with 50 residual blocks (169 layers) has approximately 25 million parameters.

Recently, an even deeper network, DenseNet (Huang *et al.*, 2017), has addressed the vanishing gradient and feature redundancy issues by using dense blocks. Vanishing gradient results from the fact that, since weight update is based on the propagation of the error derivatives' from the subsequent layer, the gradient tends to zero in very deep networks, hindering training. DenseNet introduces short connections between convolutional layers, with the activations of all preceding layers being used as inputs for a given layer. This promotes feature re-use, allowing to decrease the number of parameters (20 million parameters for a DenseNet with 201 layers).

Despite the success of deep learning approaches for medical image analysis, their tendency to overfit and their black-box behaviour may lead researchers to report overconfident claims on the performance of their systems. Indeed, as stated in Abràmoff *et al.* (2016), CNNs may learn unwanted and unexpected associations, such as the association of certain acquisition details to a given image label, instead of the intended pathological features, for instance. Also, in the BACH challenge, participants reported their performance having only access to the training set, prior to the final evaluation on a test set (hidden labels), and these showed to be almost consistently overconfident, with the results having dropped on the test set (Aresta *et al.*, 2019c). This highlights the need to properly design the experiments and to maintain a truly independent set of data for reporting a trustworthy algorithm's performance.

In sum, despite the rapid development of deep learning research for healthcare applications, and the very promising performance of current techniques, the incorporation of these systems in real-world healthcare environments is limited by several factors. Among these, we have: regulation and ethics, data availability and variability and technology issues related to machine learning solutions (Combalia *et al.*, 2020). Among the last ones, we can find essential aspects such as the fitting of confounders, interpretability or explainability of the algorithm's decisions, generalization to different data populations and estimation of reliable measures of the model's confidence or uncertainty. Most of the deep learning solutions produce deterministic outputs without quantifying the uncertainty in the predictions (Combalia *et al.*, 2020). This is of special relevance in the clinical practice since unexplained decisions, not associated to a confidence level, may lead to a lack of trust of the healthcare professionals and difficulty in interpreting the diagnosis generated by the algorithm. The fact that we can know how uncertain the model is about a prediction allows to establish a priority list, highlighting the cases that are more complex and thus would require a closer examination or second opinion (Leibig *et al.*, 2017).

2.2.2 Uncertainty of the prediction

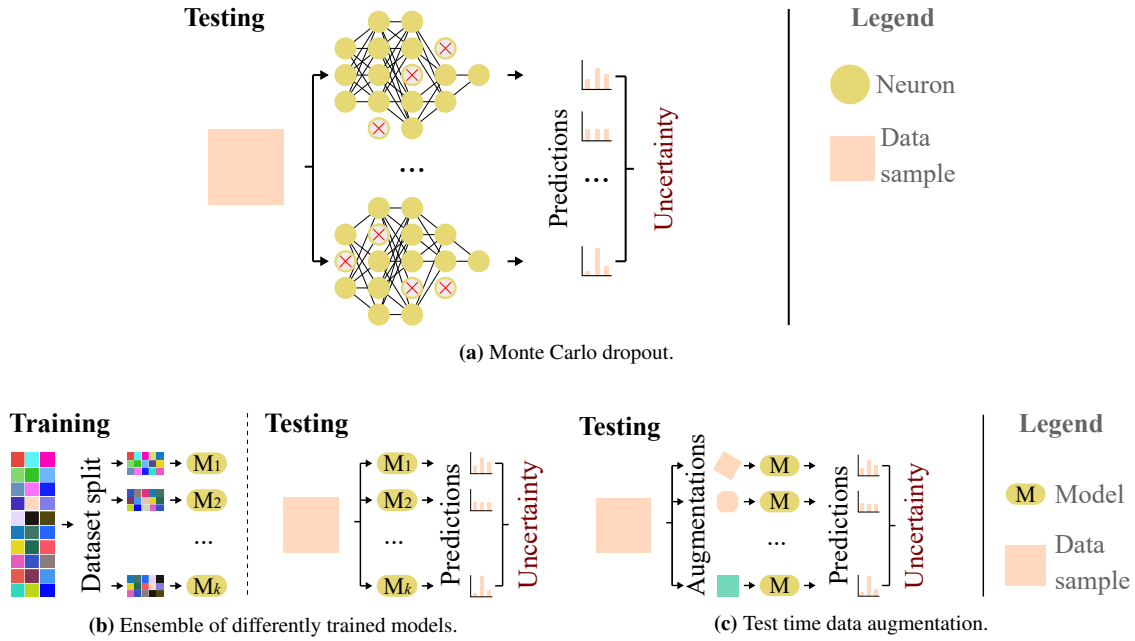
The predictive uncertainty can be divided in two types: epistemic and aleatoric uncertainty (Kendall and Gal, 2017). Epistemic uncertainty is caused by the uncertainty is the model's parameters, arising from the limited amount of training data (Laves *et al.*, 2020). It accounts for our ignorance about the model's parameters, and can be explained away given enough data (Kendall and Gal, 2017). On the other hand, aleatoric uncertainty derives directly from the data, due to noise in the observations, constituting the input-dependent uncertainty (Combalia *et al.*, 2020). While epistemic uncertainty may vanish in the case of no parameters ignorance, aleatoric uncertainty remains, due to sensor or motion noise, for instance, and cannot be explained away by capturing more data (Kendall and Gal, 2017; Ayhan and Berens, 2018).

Uncertainty in neural networks can be modeled by Bayesian analysis (Combalia *et al.*, 2020), with epistemic uncertainty being formalized as probability distributions over the model parameters, and the aleatoric uncertainty as probability distributions over the model outputs (Kendall and Gal, 2017). For epistemic uncertainty, a prior distribution is placed over the model's parameters, for instance a Gaussian prior distribution, and the variation of these weights given some input data is analysed. This model is usually called a Bayesian neural network (BNN). In these the deterministic weight parameters are replaced with distributions over these parameters (Kendall and Gal, 2017). Bayesian inference is used for computing a posterior probability over the weights of the model for a given training data and the uncertainty is estimated from the predictive distribution of a test sample (Combalia *et al.*, 2020).

BNNs are easy to formulate but the exact Bayesian inference is computationally intractable in several real-world applications (Gal and Ghahramani, 2016). Due to this, several approximations have been proposed (Lakshminarayanan *et al.*, 2017), such as the Laplace approximation (MacKay, 1992), Markov chain Monte Carlo (Neal, 1996), variational Bayesian methods (Graves, 2011), and assumed density filtering (Hernández-Lobato and Adams, 2015).

Dropout has been commonly used for performing approximate Bayesian inference (Fig. 2.2a). This method, often referred to as Monte Carlo (MC) dropout, was proposed by Gal and Ghahramani (2016) and consists in turning dropout on at test time, performing a given number of predictions for a test case, and based on these predictions obtain a proxy for the true predictive posterior. Performing multiple predictions on the same input image can be interpreted as having an ensemble of highly similar models, and thus disagreement between the inferences is indicative of the model uncertainty (Gal and Ghahramani, 2016). The method is very simple and efficient, and was successfully applied in a VGG-like CNN for DR and referable DR detection by Leibig *et al.* (2017). The authors also show that uncertainty-aware decision referral can improve the diagnosis.

Another common approach for approximating the Bayesian inference is the ensemble of differently trained networks, as performed by Lakshminarayanan *et al.* (2017), in which the predictions of multiple models are averaged to capture model uncertainty (Fig. 2.2b). The authors combined the ensemble approach with adversarial training (Goodfellow *et al.*, 2015) for robustness to out-of-distribution samples. Indeed, the production of overconfident predictions on out-



(a) Monte Carlo dropout.

(b) Ensemble of differently trained models.

(c) Test time data augmentation.

Figure 2.2: Examples of uncertainty estimation schemes.

of-distribution samples is a known challenge for reliable deployment of deep learning solutions in the real-world (Lakshminarayanan *et al.*, 2017). Ideally, predictions should be associated with higher uncertainties when test samples differ significantly from the training data.

When modeling simultaneously the two types of uncertainties, the explanation is dominated by the aleatoric uncertainty, being the contribution of the epistemic component marginal (Ayhan and Berens, 2018). It is thus more effective to address the aleatoric uncertainty, given that the epistemic one can be explained away. However, as stated in Kendall and Gal (2017), modeling aleatoric uncertainty alone comes at a cost, since out-of-data samples (identifiable with epistemic uncertainty), cannot be identified with aleatoric uncertainty alone. In Kendall and Gal (2017), the authors showed that epistemic uncertainty was indeed higher for test samples that are very different from the training samples, whereas aleatoric uncertainty remained approximately constant. Experiments also showed that the epistemic uncertainty decreased as the training set size increased.

Ayhan and Berens (2018) estimated aleatoric uncertainty based on test time data augmentation (Fig. 2.2c). A given number of augmented examples are generated per test sample and fed to a NN to obtain a predictive distribution, afterwards used for evaluating the uncertainty in the prediction. They focus on traditional data augmentation techniques, such as simple geometric and color transformations (random crop and resize, brightness, hue, saturation, and contrast adjustments, horizontal and vertical flips, and rotation).

In a more recent work, Filos *et al.* (2019) built on top of the work of Leibig *et al.* (2017). They tested different methods for uncertainty estimation, evaluating them for the binary task of referable DR detection: mean-field variational inference (Peterson, 1987; Graves, 2011), MC dropout (Gal and Ghahramani, 2016), model ensembles (Lakshminarayanan *et al.*, 2017), ensembles of MC

dropout models, and the sigmoid output probabilities (baseline). Regarding the sigmoid output baseline, it is assumed that the model is more confident the closer to 1 or 0 its output is. It is the simpler way for estimating the uncertainty, but only quantifies aleatoric uncertainty, not capturing epistemic uncertainty, *i.e.*, about the model's knowledge. The ensembles of MC dropout models was the approach which performed better in the experiments to assess the performance of the classifier in relation with the uncertainty of the prediction.

2.2.3 Explainability of the decision

Several frameworks have been proposed for visualizing/interpreting what NNs have learnt. Below we briefly describe the most common ones.

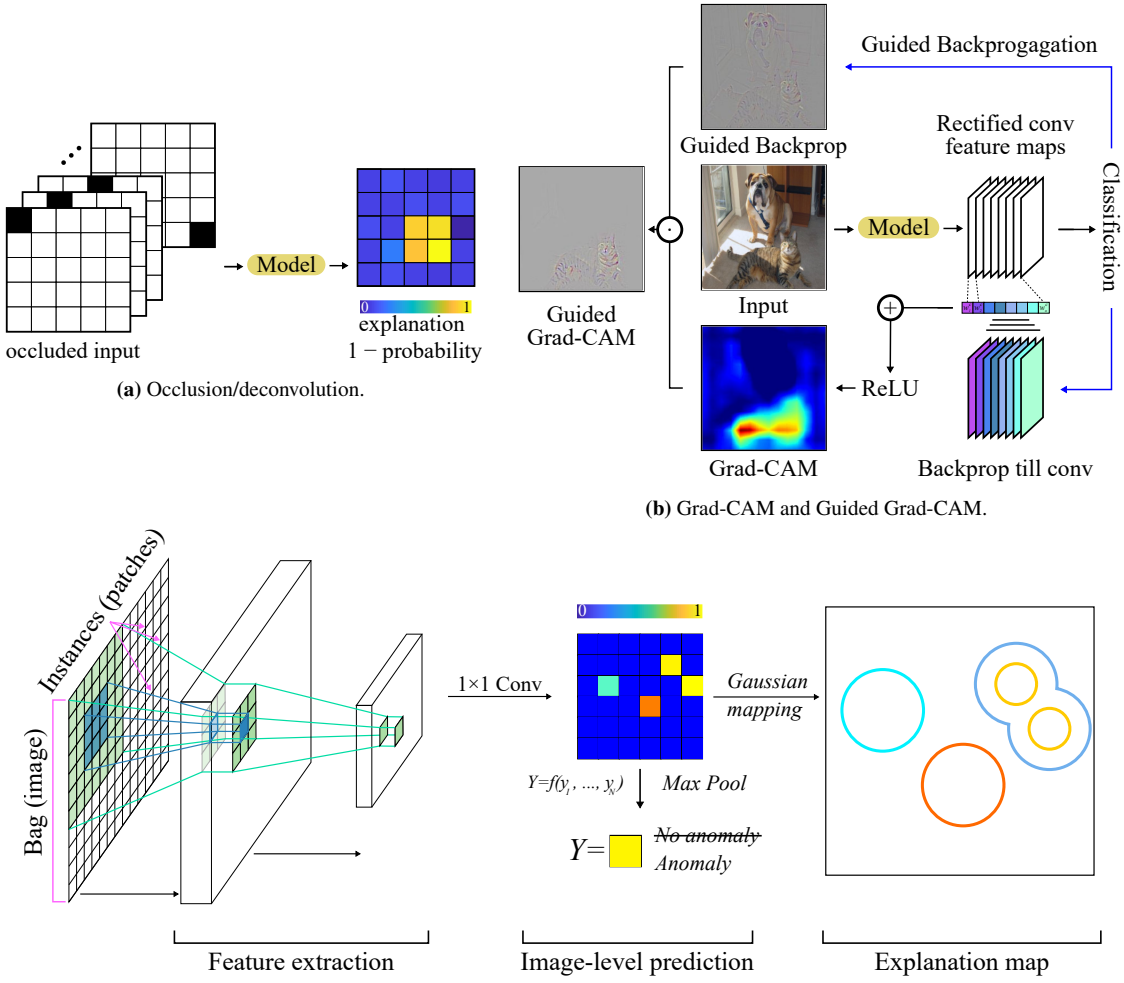
Sensitivity analysis Sensitivity analysis was used in Simonyan *et al.* (2014) to compute saliency maps of images classified by NNs. The sensitivity of a pixel (computed using the norm of partial derivatives) measures how small changes in the pixel value affect the network output in a local manner (Samek *et al.*, 2016). The main idea behind the guided backpropagation is the interpretation of the network's gradients with respect to the input image. However, given the definition of the sensitivity, in which the norm of the derivatives is used, the direction of the change is lost.

Guided backpropagation Springenberg *et al.* (2015) suggested to modify the backward pass in a manner that the negative gradients are set to zero in each layer allowing that only positive gradients flow backward through the network and thus only pixels associated with large magnitude positive gradients are highlighted (Wickstrøm *et al.*, 2020). The motivation behind this is that the high magnitude negative gradients correspond to pixels the network wants to suppress, whereas the high magnitude positive gradients are the ones that have high importance in the network's decision.

Deconvolution The deconvolution, or occlusion, method was proposed in Zeiler and Fergus (2014), and, similarly to the backpropagation method, it aims to map the activations from the network's output back to the pixel space. In this technique, patches are occluded in the input image and the new image is run through the network. A decrease of performance suggests that the patch that was occluded is relevant for the classification (Fig. 2.3a).

Layer-wise relevance propagation (LRP) LRP (Bach *et al.*, 2015), which also relies on back-propragated quantities, decomposes the network's decision into pixel-wise relevances, indicating the contribution of an input pixel to the overall classification. It is derived from a layer-wise conservation principle (Bach *et al.*, 2015), which forces the propagated quantity to be preserved between neurons of two adjacent layers (Wickstrøm *et al.*, 2020).

Integrated gradients In the integrated gradients technique (Sundararajan *et al.*, 2017), the explanation maps measure the contribution on each input pixel by computing the average gradient



(a) Occlusion/deconvolution.

(b) Grad-CAM and Guided Grad-CAM.

(c) Multiple Instance Learning.

Figure 2.3: Examples of explanation map generation techniques.

while the input varies linearly from a baseline image, which is typically an image of zeros, to their current value (Sayres *et al.*, 2019).

Grad-CAM More recently, Gradient-weighted Class Activation Mapping (Grad-CAM) has been proposed by Selvaraju *et al.* (2017). As depicted in Fig. 2.3b, given an input image and a class of interest (*e.g.*, cat), the image is forward propagated through the model to obtain a score for that class. The gradients are set to zero for all classes except the selected class, which is set to one. This signal is then backpropagated from the output to a convolutional layer of interest, which activation maps are linearly combined to compute the coarse Grad-CAM heatmap, which represents the relevant regions for the model to make that prediction. Guided Grad-CAM (Selvaraju *et al.*, 2017) combines guided backpropagation and Grad-CAM aiming at improving the localization capability of the Grad-CAM. As shown in Fig. 2.3b, this map is obtained by pointwise multiplication of the Grad-CAM and the Guided Backpropagation heatmaps.

Multiple Instance Learning A different approach to explanation map generation relies on the Multiple Instance Learning (MIL) assumption (Costa and Campilho, 2017; Costa *et al.*, 2019). The concept is exclusive to problems that can be formulated based on the MIL framework. For instance, in a binary problem, considering the image as a *bag* and each image patch as an instance, the *bag* is positive if it contains at least one positive instance. As thus, the bag is negative only if all of its instances are negative. This assumption can be implemented in a CNN architecture by a max pooling over a feature map containing all patch predictions, outputting the image-level prediction (Fig. 2.3c). That map will highlight image patches only when a positive prediction occurs, thus consisting in a way of explaining the network's decision.

Note that these methods, as they highlight features in the input space that contribute to the prediction, allow to perform weakly-supervised detection of objects (Gonzalez-Gonzalo *et al.*, 2020).

2.3 Methods for DR grading

As referred in Chapter 1, this thesis mainly focuses on the development of novel and explainable deep learning-based methodologies to help specialists in the DR grading process. As shown in Section 2.1, there are several contributions in the literature for DR diagnosis, including grading, not all resorting to deep learning. A brief review of the methods considered of more relevance for this thesis is presented in this section. This does not constitute a exhaustive review of the published studies for DR diagnosis, but rather a contextualization on the key approaches on the topic, focusing on the last decade.

Several methods have been proposed for DR assessment using eye fundus images, targeting either DR detection, referable DR detection or DR grading in the different severity levels. DR detection consists in assessing if the image is normal, *i.e.*, has no signs of DR, or pathological. Referable DR is commonly assumed as moderate NPDR or higher. As the topic of the PhD thesis is DR grading, this chapter focuses mainly on the algorithms described in the literature for this purpose. We also show relevant studies that address DR and referable DR detection.

As referred in Chapter 1, DR grading consists in attributing a DR severity level to an image, usually considering the following classes: 1) normal image (no DR); 2) mild NPDR; 3) moderate NPDR; 4) severe NPDR and 5) PDR. The differentiation of these grades relies on the observation of the typical DR lesions, such as: 1) microaneurisms; 2) hemorrhages; 3) hard exudates; 4) cotton wool spots and 5) neovessels. Depending on the type of grading performed, different lesions may be of interest. For instance, when considering PDR, NVs should be involved (directly or indirectly) in the classification.

Methods for DR assessment can be classified considering different characteristics. One is the number and type of grades retrieved by the system. Some studies consider only NPDR classes, whereas others also consider PDR. Within the NPDR cases, some studies grade DR in two or three classes (mild, moderate or severe). Alternatively, algorithms can be classified accordingly to the main type of methodology (Fig. 2.4), namely on : 1) lesion-based, and 2) non lesion-based

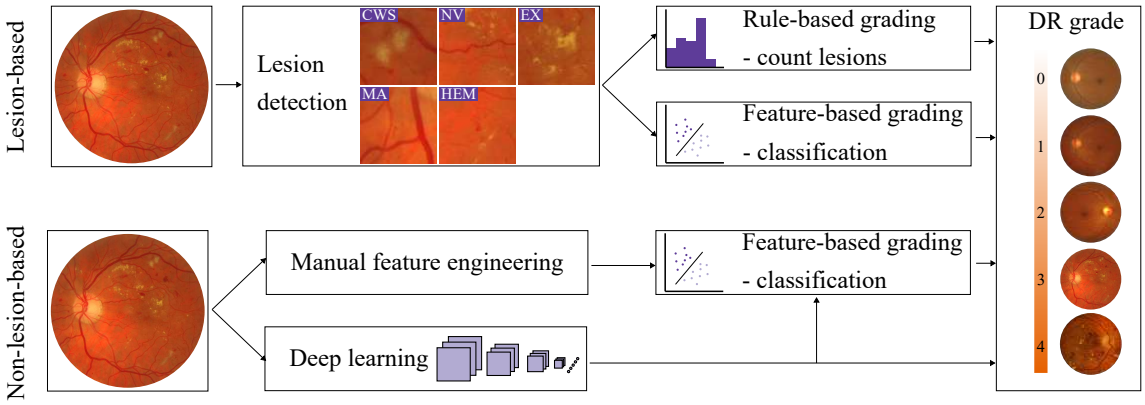


Figure 2.4: General approaches for diabetic retinopathy (DR) grading. In the case of lesion-based methods, the retinal image is analyzed for the detection of typical DR lesions, depending on the grades of DR considered. Some works then classify the image using rule-based grading (for instance, by counting the number of lesions of each type), or by using classifiers trained to extract the image-level grade, for example. In non-lesion-based approaches, features are typically extracted from the image to train a classifier that outputs the grade (classical approach). Alternatively, deep learning-based methods can be used, which automatically learn the relevant features.

methods. The first scheme relies on the detection of typical DR lesions, based on which the image grade is determined. Contrarily, methods that fit in the second scheme perform grading in a way that does not require previous lesion detection or segmentation. Note that even if indirectly, the presence of the lesions is certainly the determining factor for the prediction.

Table A.1 (Appendix A) presents a brief summary of some of the more relevant state-of-the-art methods for DR assessment, consisting on the main methodology, the datasets used for validating the approaches and the main results. The most representative studies using both non deep learning and deep learning approaches will be detailed in the remaining of this section.

2.3.1 Lesion-based methods

Lesion-based methods directly account for the typical DR lesions to grade eye fundus images, involving most of the time the use of image-analysis methodologies for detection or segmentation of the lesions.

Methods for DR lesion detection The most common framework for lesion detection start with image preprocessing and blood vessels and OD detection to avoid confusion with red and bright lesions, respectively. This preprocessing often includes intensity normalization (Zhang *et al.*, 2014; Seoud *et al.*, 2016), noise removal (Bhaskaranand *et al.*, 2016; Wu *et al.*, 2017) and illumination correction and contrast enhancement techniques (Mookiah *et al.*, 2013; Mahendran and Dhanasekaran, 2015). This is followed by lesion candidate detection, achieved by methods based on mathematical morphology (Zhang *et al.*, 2014; Amin *et al.*, 2017), filter responses (Akram *et al.*, 2014; Kar and Maity, 2018), pixel-based classification (Niemeijer *et al.*, 2007; Sánchez *et al.*, 2011), among others. The features extracted in the next phase are often related with color, size, shape, intensity and statistics (Seoud *et al.*, 2015; Dai *et al.*, 2016; Wu *et al.*, 2017), as well

as contextual features (Zhang *et al.*, 2014). Afterwards, these features are used for lesion classification, using for instance random forests (Gupta *et al.*, 2017; Seoud *et al.*, 2015), k-nearest neighbors (kNN) (Wu *et al.*, 2017; Niemeijer *et al.*, 2005), support vector machine (SVM) (Amin *et al.*, 2017; Roychowdhury *et al.*, 2014), Gaussian mixture model (GMM) (Roychowdhury *et al.*, 2014), naive Bayes (Amin *et al.*, 2017) or hybrid classifiers (Akram *et al.*, 2014).

Deep learning-based approaches have also been proposed for DR lesion detection (Abràmoff *et al.*, 2016; Prentasic and Loncaric, 2016; Quellec *et al.*, 2017). Alternatively, some works do not rely on candidate detection plus feature extraction and classification nor on deep learning, but rather perform candidate detection and post-processing to find lesions (Reza *et al.*, 2011; Nayak *et al.*, 2008; Mahendran and Dhanasekaran, 2015).

DR and referable DR detection Several studies consider only normality detection. In these, the goal is to distinguish between normal and DR-affected images (of any grade) (Seoud *et al.*, 2016; Sánchez *et al.*, 2011; Antal and Hajdu, 2012). Some of them focus on the detection of referable DR. A case is generally considered as referable if the DR grade is greater than mild, *i.e.*, moderate or worse (Abràmoff *et al.*, 2010; Bhaskaranand *et al.*, 2016; Seoud *et al.*, 2016). The latter is the main goal of DR screening systems, *i.e.*, is to identify patients who need referral to a specialist since they have high risk of vision loss (Bhaskaranand *et al.*, 2016).

Generally, methods that rely on lesion detection for DR assessment start with initial candidate detection (often after some preprocessing steps), followed by feature extraction and classification for lesion identification. Then, the presence and quantity of lesions of different types can then be assessed to determine the DR presence/stage. For instance, in normality detection studies it is common to rely only on red lesion detection (Seoud *et al.*, 2016), and even only on MAs (Antal and Hajdu, 2012). This makes sense since the first unequivocal signs of DR are red lesions (Niemeijer *et al.*, 2005), being the very first sign the presence of MAs. However, the detection of bright lesions has also been used, along with red lesion detection (Sánchez *et al.*, 2011). In this work a kNN classifier is trained with features extracted from detected candidate lesions (mainly related with the probabilities of these being true lesions), which by its turn outputs the probability of the image having DR. In Seoud *et al.* (2016), an image is attributed a probability of having DR equal to the maximum of all probabilities attributed to the detected lesions. If the image probability is higher than a established threshold then it is considered as presenting DR. Alternatively, in other studies (Antal and Hajdu, 2012), the system decides that an image has DR if at least a lesion (*e.g.* MA) has been detected in that image.

In referable DR detection, there is the need to distinguish cases of moderate DR or higher from normal or mild DR. In this sense, it is common that more lesion types are taken into account in the studies that address referable DR detection. In fact, in Bhaskaranand *et al.* (2016), MAs, HEMs, EXs, CWSs and NVs are considered. A supervised learning ensemble classifier determines if the patient is referable or not based on the detected lesions. By its turn, Abràmoff *et al.* (2016) takes advantage of deep learning to detect each type of lesion. The CNN-based detectors have an architecture inspired by the Alexnet (Krizhevsky *et al.*, 2012) and the VGG (Simonyan and

Table 2.1: Diabetic retinopathy severity scale used in Akram *et al.* (2014). NPDR: Non-proliferative DR; PDR: proliferative DR; #: number of.

Grade	Description
R0 - normal	no lesion found
R1 - Mild NPDR	number of MAs satisfies $0 <$ #MAs ≤ 5
R2 - Moderate NPDR	few MAs, HEMs or EXs are found, satisfying $5 <$ #MAs ≤ 10, $0 <$ #HEMs, #EXs ≤ 5
R3 - Severe NPDR	higher number of lesions than in moderate grade

Zisserman, 2015). Besides pathological regions, common lesion confounders are also included in the training set, such as the ones resulting from dust on the camera's lens, specular reflections or blood vessels (Quellec *et al.*, 2011). Each CNN outputs the likelihood that a detection is a real lesion, and, the feature vectors formed by these likelihoods are the fed to random forests, which return the likelihood that a patient has to be referred. The performance of this method is compared with a previous method from the authors (IDX-DR) (Abràmoff *et al.*, 2013), and results showed that the deep learning-based method had clearly better performance.

NPDR grading NPDR image grading can be performed considering three levels: mild, moderate and severe (Akram *et al.*, 2014). The detection of other relevant lesions (besides MA) is of interest for more complex gradings (Zhang *et al.*, 2014; Sánchez *et al.*, 2011). For instance, MAs, HEMs and EXs have been considered for NPDR grading in the three above categories by Akram *et al.* (2014). In this work, after classifying the lesions in their types, the retinal image is graded according to the type and count of lesions extracted, based on the rules from Table 2.1. This method fits in the category of rule-based grading, since the final image grade is determined based on a set of rules related with the count and type of lesions. Roychowdhury *et al.* (2014) also perform grading in the three NPDRs classes, but consider additionally CWSs. After lesion identification, grading is performed based on lesion counting considering the DR criteria used for the Messidor dataset annotation (Decenciere *et al.*, 2014; Pires *et al.*, 2017) (Table 2.2).

This criteria considers the number of MAs and HEMs, but not the bright lesion count. Some studies have shown that clinically relevant lesion combination for DR severity assessment relied more on accurate estimation of the number of red lesions than of bright lesions (Roychowdhury *et al.*, 2014). In fact, some works only perform red lesion detection for NPDR grading in the

Table 2.2: Diabetic retinopathy severity scale from Messidor dataset annotation. NPDR: Non-proliferative DR; PDR: proliferative DR; #: number of.

Grade	Description
R0 - normal	#MAs = 0 and #HEMs = 0
R1 -	$0 <$ #MAs ≤ 5 and #HEMs = 0
R2 -	$5 <$ #MAs ≤ 15 or $0 <$ #HEMs ≤ 5
R3 -	#MAs ≥ 15 or #HEMs ≥ 5 or #NVs ≥ 1

Table 2.3: International clinical Diabetic retinopathy (DR) disease severity scale. NPDR: Non-proliferative DR; PDR: proliferative DR.

Grade		Description
R0 -	No DR	No visible sign of abnormalities
R1 -	Mild NPDR	Presence of microaneurysms only
R2 -	Moderate NPDR	More than just microaneurysms but less than severe NPDR
R3 -	Severe NPDR	Any of the following: - >20 intraretinal hemorrhages, in each of 4 quadrants - Definite venous beading in 2/+ quadrants - Proeminent intraretinal microvascular abnormalities in 1/+ quadrant - no signs of PDR
R4 -	PDR	One or more of the following: - Neovascularization - Vitreous/pre-retinal hemorrhage

three severity classes (Seoud *et al.*, 2015; Saleh and Eswaran, 2012). In the case of Saleh and Eswaran (2012), MAs and HEMs are detected, and rule-based NPDR grading is performed based on international clinical DR disease severity scale (Wilkinson *et al.*, 2003) (Table 2.3).

Contrarily to the above mentioned methods, Seoud *et al.* (2015) perform grading in the three NPDR stages based on red lesion (MAs and HEM) detection but use a random forest classifier to determine the grade instead of using rules. This classifier is trained with features extracted from the previously segmented lesions, which are related with location, size and probabilities.

Other researches consider fewer levels of severity. For instance, Mahendran and Dhanasekaran (2015) consider only moderate and severe NPDR levels. After EXs segmentation, features based on the gray level co-occurence matrix (GLCM) (contrast, correlation, energy, homogeneity, etc.) are used for training classifiers to grade the image. Both SVM and probabilistic neural network (PNN) classifiers are tested. Please note that all the referred works also consider the normal class in their classification (besides the NPDR levels).

PDR detection PDR recognition relies mainly on the detection of NVs. Features such as blood vessel area, perimeter (Yun *et al.*, 2008), and number of vessel bifurcations (Mookiah *et al.*, 2013) have thus been used for PDR detection.

Some studies classify images in three categories: no signs of DR, NPDR or PDR. Mookiah *et al.* (2013) trained classifiers with features extracted from EXs and vessel segmentations, such as EX area, blood vessel area, vessel bifurcation points count, along with local binary patterns (LBP)-based texture features. Different classifiers are tested and compared: PNN, decision trees and SVM. Similarly, Nayak *et al.* (2008) trained a NN with textures extracted from EXs and blood vessels, including EX area and blood vessel area.

Differently, Gupta *et al.* (2017) focus on PDR detection alone, performing local characterization of NVs. Patches extracted from the image are characterized in terms of texture and vessel-based characteristics, and random forests are trained to retrieve a neovascularity score for each

patch. At the image level, an image is considered to have PDR if at least a patch has a neovascularity level greater than a given pre-established threshold.

Some works consider all the severity grades. Roy *et al.* (2017) grades images as normal, mild, moderate or severe NPDR or PDR. Although relying on deep learning, their method also involves dictionary-based approaches. In this sense, the authors claim to combine the advantages of image-based CNN methods and pathology-specific CNNs, while surpassing their drawbacks. Typically, image-based CNNs require the downsample of the image due to memory constraints, and thus have difficulty on detecting small lesions such as MAs. By their turn, lesion-based CNNs usually require a large amount of annotated data, which is difficult to obtain. The proposed method produces discriminative and generative pathology histograms, which are combined with features extracted from a CNN (fully-connected layer activations). A random forest classifier is trained using this hybrid feature representation of the image, with the purpose of grading DR.

2.3.2 Non-lesion-based methods

Non-lesion-based methods employ strategies other than lesion detection and lesion-based grading, rather than relying on the presence of DR lesions in an indirect way.

DR and referable DR detection The majority of the ongoing research is on DR and referable DR detection. Casanova *et al.* (2014) tried random forest and logistic regression classifiers for the purpose of DR detection. More recently, Gargeya and Leng (2017) applied deep learning for the same purpose, making use of deep residual nets for feature extraction. The entire downsampled retinal images are given as input to the net. Features are extracted from the global average pooling layer, and used along with metadata features to train a decision tree-based classifier. The use of a residual CNN allowed the authors to train deep nets with many parameters, due to the summations between the convolutional layers.

Gulshan *et al.* (2016) proposed ensembles of CNNs to detect referable DR in retinal images. Similarly to other works (Quellec *et al.*, 2017; Gargeya and Leng, 2017), the full downsampled retinal images are used for training the CNNs. Quellec *et al.* (2017) also used CNNs for referable DR detection. Besides retrieving binary decisions at the image-level, the authors also focused on finding which local patterns analyzed by the CNNs lead to the image-level decisions. For this purpose, they create heatmaps showing which pixels have more importance in the final predictions. Note that the authors propose a technique for detection of DR lesions that does not require manual annotations/segmentations of the lesions as ground truth. The work of Costa *et al.* (2019) builds on the Multiple-Instance-Learning assumption to propose a deep learning framework to predict the presence of DR and also pinpoint relevant regions for the diagnosis, also being trained solely on image-level labels.

Table 2.4: Modified Davis grading used in Takahashi *et al.* (2017) for the DMR dataset annotation.

Grade	Description
No DR	No signs of diabetic retinopathy
SDR	Microaneurysm, retinal hemorrhage, hard exudate, retinal edema, and more than 3 small soft exudates
PPDR	Soft exudate, varicose veins, intraretinal microvascular abnormality, and non-perfusion area over one disc area
PDR	Neovascularization, pre-retinal hemorrhage, vitreous hemorrhage, fibrovascular proliferative membrane, and tractional retinal detachment

DR grading Fadzil *et al.* (2011) used a Gaussian Bayes classifier trained with the area of the foveal avascular zone (FAZ) to grade NPDR. It has been reported that the FAZ (capillary-free zone in the macula) area increases in DR cases.

Acharya *et al.* (2012) graded DR in NPDR and PDR resorting to texture features (homogeneity, correlation, short run emphasis, long run emphasis and run percentage) for training a SVM classifier. Recently, Takahashi *et al.* (2017) proposed a deep learning based method to stage DR in the modified Davis stages: simple DR (SDR), pre-proliferative DR (PPDR) and proliferative DR (PDR), using a wider retinal area. The authors state that for DR grading there is need for more than one fundus image (contrarily to DR detection), being required four or more. This is true since the fundus photograph taken at 45° shows only the most disease-prone areas, whereas to see the entire retina an angle of 230° would be required. Because of that, they use 45° color fundus images taken of four fields, *i.e.*, quadrants, in each eye.

Deep learning-based methods have more recently emerged for solving the task of DR grading in all five stages. (Krause *et al.*, 2018) built on the previous work of Gulshan *et al.* (2016), extending it to perform grading in five stages, increasing the amount of training data and input image size, among others. Other works proposed CNNs for DR grading, with some focusing on the ordinal nature of the grading problem, proposing loss functions that consider the distance between the grades (de la Torre *et al.*, 2018; Zhao *et al.*, 2019; Galdran *et al.*, 2020b). Gonzalez-Gonzalo *et al.* (2020) also applied CNNs for DR grading, additionally proposing a framework for interpretability of the decisions via an iterative procedure aiming at improving the results of existent methods for the generation of explanation maps.

2.3.3 Relevant research contributions

In this subsection, the studies that are considered of more relevance to this thesis are further analysed. These consists mostly of recent works (from 2016 or after), based on deep learning, including some which already consider the need for explainability of their predictions.

Gulshan *et al.* (2016) proposed ensembles of CNNs to detect referable DR in retinal images. The full downsampled retinal images (299 pixels wide) are used for training the CNNs, which consists of an Inception-v3 (Szegedy *et al.*, 2016) pre-trained on the ImageNet dataset (Russakovsky

et al., 2015). The algorithm was trained with 128 175 images retrieved from the EyePACS in the United States and three eye hospitals in India. All images were graded by 3 to 7 ophthalmologists from a panel of 54 US specialists in the following stages: none, mild, moderate, severe and proliferative DR, accordingly to the International Clinical Diabetic Retinopathy Scale (American Academy of Ophthalmology, 2002) (Table 2.3). As referred by the authors, although the algorithm does not explicitly detect lesions, it likely recognizes them through the learned local features.

An attempt to reproduce the results from the work of Gulshan *et al.* (2016) resorting to publicly available datasets was performed in Voets *et al.* (2019). The authors re-implemented the method from the original study, and trained the network using the Kaggle dataset. This data was annotated by a single grader, whereas in Gulshan *et al.* (2016) several ophthalmologists were involved in the annotation. The authors failed to reproduce the original study's results using only public data. They attribute this likely to the use of single grade per image or to the different data. This work shows the challenges of reproducing published results, either by lack of disclosing of the data or the source code/training details. It also hints at the very large influence of the type, amount, and mainly the quality of training data.

The work of Krause *et al.* (2018) was built upon the network proposed by Gulshan *et al.* (2016). The authors finetuned an Inception-v4 (Szegedy *et al.*, 2017), instead of the Inception-v3, and predicted five class grading for DR, instead of the binary prediction of the first study. Further, the model is now trained with a much larger set of images (more than 1.6 million images) and of higher resolution (779×779 instead of 299×299 pixels). This increase in the number of pixels aims at avoiding the elimination of small lesions such as MAs. To account for data imbalance, the authors apply a cascade of thresholds on the output probabilities of the model for each DR grade to obtain the final image classification. These thresholds are sequentially applied in a severity descending order (*i.e.*, starting on the proliferative DR), to ensure that a minimum grade-wise sensitivity is achieved.

Instead, de la Torre *et al.* (2018) considered DR grading as an ordinal classification problem and thus aimed at reducing the distance between predicted and true labels of an image. This is based on the fact that in an ordinal problem it is less wrong to predict a class that is closer to the ground truth one than predicting one that is more distant. For that, they proposed a quadratic weighted Kappa (QWK)-based loss function, which allowed to improve the performance of a DR grading network comparing to training with the cross-entropy loss. The cross-entropy loss penalizes all wrong predictions in the same manner, whereas the QWK-based loss accounts for the distance between the classes and penalizes predictions accordingly to the quadratic distance. The authors tested different input images sizes (squares with side of 128, 256, 384 and 512 pixels), and the highest resolution yielded the best performance. To overcome the high class imbalance of the dataset, the data is artificially balanced using data augmentation, oversampling the least represented classes.

Also considering the distance between the grades, Zhao *et al.* (2019) used a weighted softmax with a distance-based weight function. The authors propose a new DL architecture, BiRA-Net, combining an attention model for feature extraction and a bilinear model for more fine-grained

classification. A ResNet (He *et al.*, 2016) is used for feature extraction, and the Attention Net is applied for focusing on the suspected area. For fine-grained classification a bilinear strategy is applied, in which two of these ResNet-Attention Net blocks are trained simultaneously for improving classification.

More recently, Galdran *et al.* (2020b) also considered the ordinal nature of the DR grading problem, by introducing a cost in the loss function that penalizes predictions accordingly to the distance between the prediction and the ground truth. Additionally, they propose to leverage existent information regarding interobserver disagreement to improve DR grading performance. A greater penalty is applied on erroneous predictions when the label is likely to be reliable and be more tolerant with wrong predictions for which the associated label is unreliable. This is based on the knowledge regarding inter-observer disagreement collected from Krause *et al.* (2018), which shows the disagreement between retinal specialists and an adjudicated consensus. The cost employed in the loss is the average of the distance-based cost and the interobserver disagreement-based cost. This cost term is added to a base loss to obtain the final loss, since, according to the authors, without this base loss the model would get stuck in local minima and classify all images in R0 and R2 (more dominant classes in the training set). Three base losses were considered: cross-entropy loss, focal loss (Lin *et al.*, 2017) and non-uniform label smoothing loss (Galdran *et al.*, 2020a), having yielded very similar results. The CNN used was the ResNeXt50 (Xie *et al.*, 2017), pre-trained on the ImageNet dataset, and the class imbalance was handled by oversampling of the minority classes.

Explainaibility of the decisions

Despite the high classification performance of several of the proposed DL systems, their black-box behavior hinders their application in the real-world. Providing explanations for DR grading is challenging due to the complexity of the task, and thus research has been mainly focused on the binary detection and referral tasks.

Quellec *et al.* (2017) focused on the interpretability of their automatic method for referable DR detection. The aim was to detect lesions or other DR biomarkers, without supervision at the lesion-level, but rather at the image-level. For this purpose, they propose to create heatmaps showing which pixels have more importance in the final predictions, being the explanation map jointly optimized with the CNN prediction during training. The method is based on the sensitivity analysis from Simonyan *et al.* (2014), which computes the partial derivatives of the image-level predictions with respect to each pixel value. These partial derivatives are backpropagated through the network to produce the attention map. Quellec *et al.* (2017) state, however, that these pixel-level visualization techniques produce heatmaps' artifacts caused by the architectures of the CNNs. The authors thus propose to reduce these artifacts by jointly optimizing the CNN predictions and the heatmaps. The heatmaps' quality is improved, *i.e.*, the artifacts are attenuated, by enhancing the heatmaps' sparsity during training. The common backpropagation method used in training had to be adapted so that it was possible to train nets that produce high quality heatmaps, since these

heatmaps depend on backpropagated quantities. The CNN architecture employed was based on the second best solution from the Kaggle competition*.

Differently, Costa *et al.* (2019) proposed the generation of an explanation map via a pre-trained network coupled with a Multiple Instance Learning (MIL)-based approach for DR detection. The network is cropped at early layers to reduce the model's receptive field, and the classification is performed by a 1×1 convolution and global max-pooling. This is based on the assumption that the presence of one patch with DR signs is enough to label the entire image as unhealthy. The output of this convolution is a map in which each element indicates the malignancy score of a corresponding patch of the input image, thus providing an explanation of why that image-level label was predicted. Three networks were tested: VGG-16 (Simonyan and Zisserman, 2015), ResNet50 (He *et al.*, 2016) and Inception-v3 (Szegedy *et al.*, 2016), with the last yielding the best DR detection performance.

Also focusing on the interpretability of their network's predictions, González-Gonzalo *et al.* (2018) applied an iterative inpainting approach that allows to identify progressively less discriminative pixels of the image for DR grading. A saliency map (Simonyan *et al.*, 2014) is computed at each step, *i.e.*, the derivative of the classification output with respect to the input image using guided back-propagation. Similarly to previously mentioned grading approaches, the problem is treated as being ordinal, with the model being optimized using a mean-square error loss. To avoid overfitting due to the high class imbalance, the classes are balanced considering the referal/non-referal sample distribution. In Gonzalez-Gonzalo *et al.* (2020), the authors assess their iterative inpaiting technique in terms of weakly-supervision lesion detection performance, and evaluate the influence in the performance of the use of different visual attribution techniques: saliency (Simonyan *et al.*, 2014), guided backpropagation (Springenberg *et al.*, 2015), integrated gradients (Sundararajan *et al.*, 2017), Grad-CAM (Selvaraju *et al.*, 2017) and Guided Grad-CAM (Selvaraju *et al.*, 2017). The authors state that their method allows to overcome the limitation of previous interpretability framework of only highlighting the most significant regions, ignoring lesions that have less influence in the network prediction but that could be important for DR understanding and grading. However, the method does not distinguish between lesions, producing a single heatmap simultaneously highlighting lesions from different grades.

It is also of great relevance to assess how these explanations maps affect the diagnostic behaviour of the health professionals in the real-world practice. Sayres *et al.* (2019) studied the influence of the algorithm from Krause *et al.* (2018) and an integrated gradients explanation in the DR grading performance of human readers. The explanation heatmaps were generated using the method from Sundararajan *et al.* (2017), which produces pixel-based maps measuring the contribution of each image pixel to a predicted DR grade. This is measured with relation to a baseline image, which in their study was a black image. The overall conclusion was that deep learning algorithms can improve the DR diagnosis performance and confidence in an assisted read setting. Despite also having increased grading time, this effect may be mitigated with experience.

*https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15617

2.3.4 Comparative overview of the existent techniques

Several methods have been proposed for DR assessment, with a significant amount of them focusing on two-class classification problems, not addressing the grading in the internationally defined severity levels. Further, some methods are validated in private or small subsets of public datasets, thus hindering proper validation and comparison. Also, even when using the same dataset, different evaluation schemes can be used (*e.g.*: crossvalidation, simple split in training and test sets, etc.). Further, external observers are sometimes involved in some studies in order to obtain annotations for the used datasets, and the annotation process and outcomes are not always clearly described. Finally, even though some authors perform DR grading they do not present all the relevant results for proper assessment of their approaches, such as confusion matrices for evaluating the performance per grade and not only the overall grading performance. Nevertheless, we analyze below the methods' performance in a broad manner.

From the results presented in Table A.1, the majority of the studies achieve high values of area under the curve (AUC), almost consistently $\geq 90\%$, for the task of DR referable and DR detection, with deep learning-based methods presenting the most promising performances (Gulshan *et al.*, 2016; Quellec *et al.*, 2017; Gargeya and Leng, 2017; Costa *et al.*, 2019). Differently from most authors, who train CNNs directly from the full images, Abràmoff *et al.* (2016) proposed the use of individual CNNs for detection of each lesion type, under the assumption that training CNNs to directly detect DR from full retinal images may lead to unanticipated associations.

Note, also, that in some of these studies a large amount of data (images and annotations) is available for training the deep learning models, being close to 100 000 images (Gulshan *et al.*, 2016; Abràmoff *et al.*, 2010; Gargeya and Leng, 2017). Indeed, Krause *et al.* (2018), who reported a quadratic weighted kappa for DR grading of 0.84, trained their network with more that 1.5 million images. Indeed, an attempt to reproduce the work of Gulshan *et al.* (2016) resorting to publicly available datasets (Voets *et al.*, 2019) failed to reproduce the original study's results, likely due to the use of single grade per image and different data, thus highlighting the importance of the type, amount, and quality of training data. When the authors do not resort to private datasets, the Kaggle DR challenge dataset is the most commonly used for training the networks (de la Torre *et al.*, 2018; Zhao *et al.*, 2019; Gonzalez-Gonzalo *et al.*, 2020; Galdran *et al.*, 2020b), likely since it has by far the largest number of images. These approaches commonly achieved kappa scores in the range 0.74-0.80 in that Kaggle test set.

Most of the works applying deep learning for DR assessment based on the entire image use downsampling to reduce computational power requirements as well as the number of parameters to optimize. Typically images are downsampled to between 300 and 512 pixels wide (Gulshan *et al.*, 2016; Gargeya and Leng, 2017; Quellec *et al.*, 2017; Costa *et al.*, 2019; Gonzalez-Gonzalo *et al.*, 2020; Galdran *et al.*, 2020b), depending on the used CNN architecture. However, a very aggressive downsampling can cause some small and more subtle lesions, such as MAs, to almost disappear. Intuitively, reducing the resolution to a very small level could cause incorrect identification of

the DR grade. However, several well performing techniques resort to 512×512 pixel images, which suggests this resolution is enough to perform proper DR assessment. On the other hand, considering the remarks made by Abràmoff *et al.* (2016), when training CNNs with the full images, the networks could be learning to associate the presence of disease with other image characteristics not related with the presence of lesions, such as position of the OD in the image.

Recently, methodologies that also focus on the importance of the explanation of the results have been proposed (Quellec *et al.*, 2017; Costa *et al.*, 2019; Gonzalez-Gonzalo *et al.*, 2020). They rely on the training of commonly used networks such as the VGG and the Inception, and implement additional features that allow to obtain explanation maps that can be shown to a clinician as the relevant regions for that classification. Most of these efforts have been made in binary decision problems, as DR and referable DR detection. However, the work of (Gonzalez-Gonzalo *et al.*, 2020) performs grading in all the five classes and proposes the generation of an explanation map, but produces a single map and not differentiates between each lesion/grade. These explanation techniques are of special importance when considering the difficulty a doctor may have in trusting blindfoldly in an algorithm's decision.

2.4 Retinal imaging

The majority of the above described methodologies associate a grade to a single fundus image of an eye. However, in the screening practice, it is common that two or more images are available per eye. Despite the Early Treatment Diabetic Retinopathy Study (ETDRS) seven standard field stereoscopic photographs remaining the gold standard for documentation of DR, they still only cover around 30% of the retina (Sun *et al.*, 2018). Advances in retinal imaging, with the ultrawide-field fundus or fluorescein angiography imaging, allow to cover up to 82% of the retinal surface (Sun *et al.*, 2018). Aditionally, advanced retinal imaging techniques such as Optical Coherence Tomography Angiography (OCT-A) allow visualization of the retinal vessels in 3D without the need of a contrast agent. Nevertheless, the commonly used publicly available datasets for DR grading are composed of one or two eye fundus images per eye, with single image-level DR label, which are described below.

2.4.1 Datasets for DR grading

In this section, the most relevant publicly available datasets for DR grading of eye fundus images are analyzed. These consist in the datasets that are more commonly used in the literature for evaluating the DR grading methods, and/or the ones that present a considerable number of images. Table 2.5 presents a summary of the main characteristics of each of these datasets. Additionally, a new private dataset is briefly described in the end of the section (the SCREEN-DR dataset).

Kaggle DR Detection This dataset was published by Kaggle[†] in 2015 in the context of a competition on DR detection. It comprises a large number of high resolution images (approximately 35 000 in the training set, 55 000 in the test set). Images were acquired under a large variety of conditions, using different types of cameras. Images are labeled by a single clinician with the respective DR grade, out of 4 severity levels: 1 - mild, 2 - moderate, 3 - severe, and 4 - proliferate DR. Level 0 stands for the absence of DR. The organizers state that this dataset is similar to any real-world dataset, in which noise is present in both images and labels. Consequently, images may contain artifacts, be out of focus and/or with inadequate exposition, aiming at promoting the development of robust and reliable algorithms that behave well in the presence of noise and across a variety of images.

Messidor This is a publicly available dataset (Decenciere *et al.*, 2014) since 2008, and is composed of 1 200 images. Images have different resolutions: 1440×960 pixels, 2240×1488 pixels and 2304×1536 pixels. Acquisition is performed by 3 ophtalmological departments, using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph at a FOV of 45°. Images are annotated by specialists, being a DR severity level/grade indicated for each image. However, no manual annotations, such as lesions position or delineation, are available. The retinopathy grading in this dataset accounts for 3 disease stages, and is given accordingly to the number of lesions, as presented in Table 2.2. Further information about this dataset is available on ADCIS, where download can be requested.

Messidor-2 The Messidor-2 public dataset (Abràmoff *et al.*, 2016)[‡] consists of color eye fundus images, one fovea-centered image per eye, of 874 subjects with diabetes, for a total of 1 748 images. Images were acquired using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic fundus camera with a 45° FOV, at 1440×960, 2240×1488, or 2304×1536 pixels (px). The ground truth (Krause *et al.*, 2018)[§] corresponds to an adjudicated consensus of three specialists. A total of 1 744 images were used, as the remaining were adjudicated as ungradable.

IDRID The Indian Diabetic Retinopathy Image Dataset (IDRiD) dataset (Porwal *et al.*, 2018) is composed of 413 images of 4288×2848 px, 50° FOV, acquired using a Kowa VX-10 alpha digital fundus camera. Images were classified accordingly to the International Clinical Diabetic Retinopathy Scale (Table 2.3). This dataset was released during the IDRID challenge (Porwal *et al.*, 2020), which was composed of different tasks, being this dataset associated with the DR and macular edema (ME) classification tasks. Other tasks included optic disc segmentation and fovea detection, and lesion segmentation.

DMR The dataset announced in Takahashi *et al.* (2017) (DMR) is composed of 9 939 posterior pole color fundus images (2720×2720 px) from 2 740 diabetic patients. Images were captured

[†] https://www.kaggle.com/c/diabetic-retinopathy-detection
[‡] http://www.adcis.net/en/third-party/messidor2/
[§] https://www.kaggle.com/google-brain/messidor2-dr-grades

Table 2.5: Publicly available datasets of color eye fundus images for Diabetic Retinopathy Grading.

Dataset	Number of DR grades	Number of annotators	Number of images	Size (pixels)	FOV (º)	Format
Kaggle DR detection	5	1	35 127	-	-	JPEG
Kaggle - test	5	1	55 000	-	-	JPEG
Messidor	4	-	1200	1440×960 2240×1488 2304×1536	45	TIF
Messidor-2	5	3 (adjudicated consensus)	1748	1440×960 2240×1488 2304×1536	45	PNG JPEG
DMR	4 (Davis scale)	1	9939	2720 × 2720	45	JPEG
IDRID	5	2 (adjudicated consensus)	413	4288×2848	50	JPEG

using a NIDEK AFC-30 fundus camera, and have a 45° FOV. One to four images may exist from a given patient. Images were graded by a modified Davis grading, described in Table 2.4.

SCREEN-DR This private dataset consists of retinal images from a Portuguese DR screening program, managed by the Portuguese North Health Administration (ARSN). A subset of 966 images was selected (SCREEN-DR dataset). These were acquired using Canon CR-2 AF nonmydriatic retinal cameras and have different resolutions (with width and height ranging approximately from 1500 px to 2600 px). Images from both left and right eyes were included, which may be centered at the macula, at the optic disc or elsewhere. Images were graded in 5 levels by a retinal specialist. For 348 of these images, pixel-wise annotations of typical DR lesions are also available.

2.5 Concluding remarks

Several methods have been proposed for DR assessment. However, a large percentage of them focus on two-class classification problems, not addressing DR grading in the internationally defined severity levels: normal (no signs of DR), mild NPDR, moderate NPDR, severe NPDR and PDR. Recently, several studies emerged presenting very promising results, but there is still room for improvement in automatic DR grading. Also, a large number of methodologies are validated in private or small subsets of public datasets, thus hindering the proper validation of the methods.

Deep learning-based techniques tend to outperform classical feature engineering approaches, since they are capable of learning and relating relevant features in a complex manner without requiring field knowledge. However, the large number of parameters to train makes the success of these models highly data reliant, both in terms of number of graded retinal images and on the quality of the annotations. Also, considering that the recovery and interpretation of the learned features is non-trivial, their behavior is most of the times hard to explain to specialists. This greatly hinders the application of deep learning-based methods for computer-aided diagnosis in

the medical centers, as it is difficult to trust a second opinion that may appear as unfounded. Nevertheless, efforts have recently started to be made to promote the ease of application of the developed deep learning-bases methods on the real-world practice, which is also a goal of this thesis.

Chapter 3

UOLO - automatic object detection and segmentation in biomedical images

This chapter is partially based on the following publication:
Araújo, T., Aresta, G., Galdran, A., Costa, P., Mendonça, A. M., *et al.* UOLO - Automatic object detection and segmentation in biomedical images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11045 LNCS:165–173, 2018a. doi: 10.1007/978-3-030-00889-5_19

In this chapter we propose UOLO, a novel framework for the simultaneous detection and segmentation of structures of interest in medical images. UOLO consists of an object segmentation module which intermediate abstract representations are processed and used as input for object detection. The resulting system is optimized simultaneously for detecting a class of objects and segmenting an optionally different class of structures. UOLO is trained on a set of bounding boxes enclosing the objects to detect, as well as pixel-wise segmentation information, when available. A new loss function is devised, taking into account whether a reference segmentation is accessible for each training image, in order to suitably backpropagate the error. We validate UOLO on the task of simultaneous optic disc detection, fovea detection, and optic disc segmentation from retinal images, achieving state-of-the-art performance on public datasets.

3.1 Introduction

Detection and segmentation of anatomical structures are central medical image analysis tasks since they allow to delimit Regions-Of-Interest (ROI), create landmarks and improve feature collection. In terms of segmentation, Deep Fully-Convolutional (FC) Neural Networks (NNs) achieve the highest performance on a variety of images and problems. Namely, U-Net (Ronneberger *et al.*, 2015) has become a reference model – its auto-encoder structure with skip connections enables the propagation from the encoding to the decoding part of the network, allowing a more robust multi-scale analysis while reducing the need for training data.

Similarly, Deep Neural Networks (DNNs) have become the technique of choice in many medical imaging detection problems. The standard approach is to use networks pre-trained on large datasets of natural images as feature extractors of a detection module. For instance, Faster-R CNN (Ren *et al.*, 2017) uses these features to identify region-of-interests (ROIs) via a specialized layer. ROIs are then pooled, rescaled and supplied to a pair of Fully-Connected NNs responsible for adjusting the size and label the bounding boxes. Alternatively, YOLOv2 (Redmon and Farhadi, 2017) avoids the use of an auxiliary ROI proposal model by directly using region-wise activations from pre-trained weights to predict coordinates and labels of ROIs.

When a ROI has been identified, the segmentation of an object contained on it becomes much easier. For this reason, the combination of detection and segmentation models into a single method is being explored. For instance, Mask-R CNN (He *et al.*, 2017) extends Faster-R CNN with the addition of FC layers after its final pooling, enabling a fine segmentation without a significant computational overhead. In this architecture, the segmentation and detection modules are decoupled, *i.e.* the segmentation part is only responsible for predicting a mask, which is then labeled class-wise by the detection module. However, despite the high performance achieved by Mask-R CNN in computer vision, its application to medical image analysis problems remains limited. This is due to the large requirement of data annotated at a pixel level, which is usually not available in medical applications.

3.1.1 Optic disc and fovea detection and segmentation

The optic disc (OD) is a bright yellow round or oval region through which the optic nerve and the vessels that supply the retina enter the eye (Yu *et al.*, 2012). The macula, by its turn, is the center of retina which is responsible for our central vision. The fovea (FV) is a small depression in the center of the macula, is darker and has no vasculature. The OD and fovea are separated approximately at a constant distance of 2.5 OD diameters (Kamble *et al.*, 2017).

The detection and segmentation of the OD and FV anatomical structures play an important role in eye fundus image analysis. The OD segmentation and FV location can be relevant for the automated diagnosis of several diseases, such as glaucoma, diabetic maculopathy, or age-related macular degeneration (AMD). Also, the OD segmentation is used in the determination of important indexes related with vascular changes, such as the arterioral-to-venular ratio (AVR). Indeed, the OD diameter is often used as a reference to measure the size or location of other anatomical and pathological structures in the retina (Al-Bander *et al.*, 2018). The OD can also be a confounder in bright lesion detection, as hard exudates and cotton wool spots. By its turn, the FV is often a red lesion confounder. Thus, the knowledge of the location of these structures may be useful for automatic DR diagnosis.

Several studies have focused on automatic OD segmentation in retinal images, which are usually divided in the following categories: template-based methods (Aquino *et al.*, 2010; Dashtbozorg *et al.*, 2015; Roychowdhury *et al.*, 2016), deformable models (Yu *et al.*, 2012), morphology-based methods (Morales *et al.*, 2013) and pixel classification methods (Cheng *et al.*, 2013). Aquino *et al.* (2010), for instance, used a template-based approach with morphological and edge detection

operations followed by Hough circle detection for segmenting the OD. Dashtbozorg *et al.* (2015), by its turn, propose the use of a multiresolution sliding band filter (SBF) (Quelhas *et al.*, 2010), a convergence filter, for the enhancement of bright circular regions. The method was developed aiming at robustness in pathological images and independence of the acquisition settings. A first SBF is applied on downsampled images to estimate the OD center, and a second SBF is applied on the high resolution images for fine OD boundary estimation. Roychowdhury *et al.* (2016) relies on a circular structuring element for morphologically reconstructing the fundus image's green channel, from which bright regions that appear near to the major blood vessels are extracted. These are classified as OD candidates or not, and the final boundary is selected based on maximization of the number of major vessel pixels and solidity.

Differently, Yu *et al.* (2012) resorted to deformable models to determine the OD boundary. OD candidates are identified using template matching, and the OD center based on vessel characteristics. The OD boundary is then found using a fast hybrid level set method, in which the curves evolves based on region and local gradient information.

Framed on the morphology-based methods, Morales *et al.* (2013) combined principal component analysis (PCA) with several morphological operators to extract the OD contour. The final segmentation is obtained via watershed-based operations. By their turn, Dai *et al.* (2017) present a method based on a sparse coding technique for OD localization, followed by an initial boundary extraction by a circular Hough transform, which evolves to the final boundary using a variational model integrating three energy terms. By their turn, Salazar-Gonzalez *et al.* (2014) start by extracting the retinal vasculature using graph cuts, based on which the OD is located. The OD segmentation is perfomed via two alternative method: Markov random field (MRF) image reconstruction and compensation factor, both use the convergence feature of the blood vessels to find the OD.

Regarding the FV detection, methods commonly rely on the knowledge of its anatomical position relatively to the OD or vascular tree. For instance, Gegundez-Arias *et al.* (2013) uses this information to search for a pixel within the foveal region, which is then used for extracting a patch containing the FV. The FV center is then found by thresholding and feature extraction. Similarly, Aquino (2014) first estimated an acceptable FV center using this anatomical knowledge, and then applies morphological processing to improve this estimation.

Other studies have focused on both tasks. For instance, Kamble *et al.* (2017) propose the fast detection of the OD and FV using 1D intensity profile analysis. The method utilizes both time and frequency domain information. The FV is detected with the help of the OD via signal valley analysis. Al-Bander *et al.* (2018), most of the studies exploited the visual appearance or anatomical features, suffering when these features are weak/invisible, for instance due to pathologies. Others use machine learning and hand-crafted feature extraction, depending on the type an quality of these features. The authors propose a multiscale sequential deep learning framework for the simultaneously detection of the FV and OD, with a first stage CNN outputting ROIs, and the second stage CNNs the refined OD and FV centers.

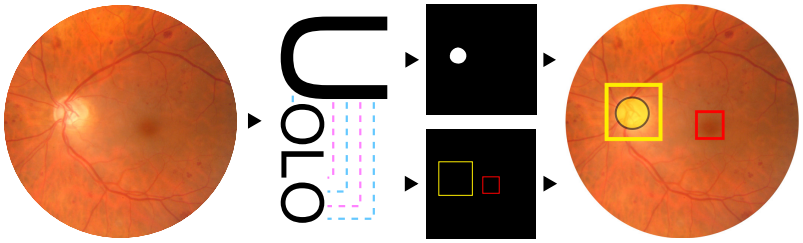


Figure 3.1: Using UOLO for fovea detection and optic disc detection and segmentation.

In this chapter, we present UOLO (Fig. 3.1), a novel architecture that performs simultaneous detection and segmentation of structures of interest in biomedical images. UOLO harvests the best of its individual detection and segmentation modules to allow robust and efficient predictions even when few training data is available. Moreover, training UOLO is simple since the entire network can be updated during back-propagation. We experimentally validate UOLO on eye fundus images for the joint task of fovea (FV) detection, optic disc (OD) detection, and OD segmentation, where we achieve state-of-the-art performance.

3.2 UOLO framework

UOLO framework for object detection and segmentation is depicted in Fig. 3.2, where the segmentation module itself is used as a feature extraction module, and serving as input for the localization module. The intuition behind this design is that the abstract representation learned by the decoding part of U-Net contains multi-scale information that can be useful not only to segment objects, but also to detect them. In addition, the class of objects that UOLO can detect is not limited to those for which segmentation ground truth is available.

3.2.1 Object segmentation module

For object segmentation we consider an adapted version of the U-Net network presented in Ronneberger *et al.* (2015). U-Net is composed of FC layers organized on an auto-encoder scheme, which allows to obtain an output of the same size of the input, thus enabling pixel-wise predictions. Skip connections between the encoding and decoding parts are used for avoiding the information loss inherent to encoding. The model's upsampling path includes a large number of feature channels with the aim of propagating the multi-scale context information to higher resolution layers. Ultimately, the segmentation prediction results from the analysis of abstract representations of the images from multiple scales, with the majority of the relevant classification information being available on the decoder portion of the network due to the skip connections. We modify the network by adding batch normalization after each convolutional layer, and replacing the pooling layers by convolutions with stride. Batch normalization (Ioffe and Szegedy, 2015) allows to increase the stability of a NN by normalizing the output of a previous activation layer taking into account the batch mean and standard deviation.

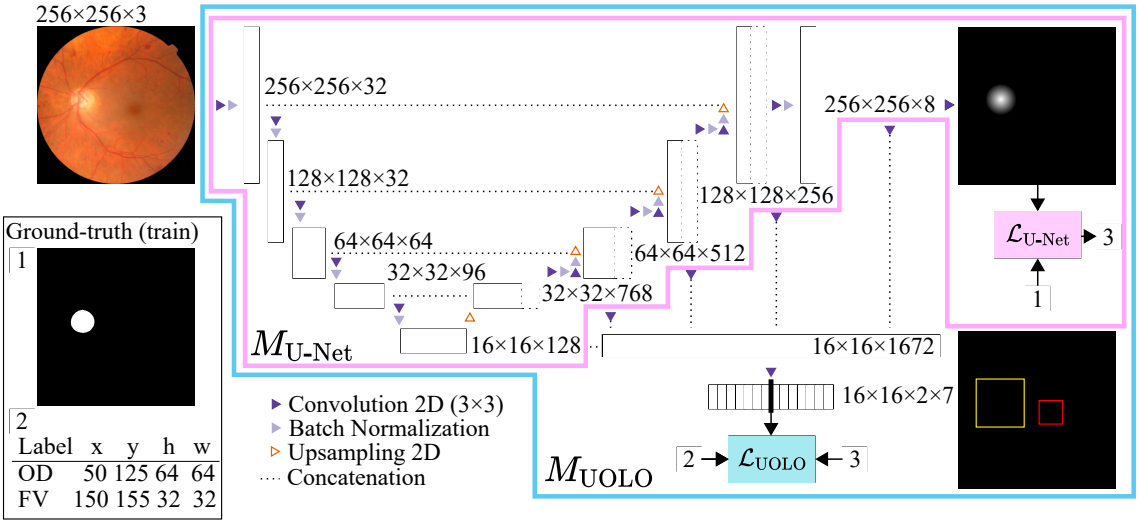


Figure 3.2: UOLO framework, nesting an U-Net responsible for segmentation and feature extraction for an YOLOv2-based detector. $M_{\text{U-Net}}$: U-net part; M_{UOLO}: full UOLO.

The soft intersection over union (IoU) is used as loss:

$$\mathscr{L}_{\text{U-Net}} = 1 - \text{IoU} = 1 - \frac{\sum I_t \circ I_p}{\sum (I_t + I_p) - \sum I_t \circ I_p}, \tag{3.1}$$

where I_t and I_p are the ground truth mask and the soft prediction mask, respectively, and $\circ$ is the Hadamard product.

3.2.2 Object detection module

For object detection we take inspiration from YOLOv2 (Redmon and Farhadi, 2017), a network composed of: 1) a DNN that extracts features from an image (F_{YOLO}); 2) a feature interpretation block that predicts both labels and bounding boxes for the objects of interest (D_{YOLO}). YOLOv2 assumes that every image's patch can contain an object of size similar to one of various template bounding boxes (or *anchors*) computed *a priori* from the objects' shape distribution in the training data.

Let the output of F_{YOLO} be a tensor F of shape $S \times S \times N$, where S is the dimension of the spatial grid and N is the number of maps. The model convolves and reshapes F into Y, a tensor of shape $S \times S \times A \times (C+5)$, where A is the number of anchors, C is the number of object classes, and 5 is the number of variables to be optimized: center coordinates x and y, width w, height h, and the confidence c (how likely is the bounding box an object) of the bounding boxes. For each anchor A_k in Y, the value of each feature map element $m_{i,j}$ is responsible for adjusting a property

Algorithm 1 Loss computation scheme of UOLO. $M_{\text{U-Net}}$: U-Net part from the UOLO model; M_{UOLO}: full UOLO model; b_{det}: batches of images with objects' bounding boxes ground truth; b_{seg}: batches of images with segmentation ground truth.

$\mathscr{L}_{\text{U-Net}} \leftarrow 1$
for each training step **do**
  $M_{\text{UOLO}} \leftarrow \texttt{train}(M_{\text{UOLO}}, b_{\text{det}}, \mathscr{L}_{\text{UOLO}})$ {train on n_{det} batches from b_{det}, backpropagating $\mathscr{L}_{\text{UOLO}}$};
  $\texttt{update}(\mathscr{L}_{\text{YOLO}})$
  $M_{\text{U-Net}} \leftarrow \texttt{train}(M_{\text{U-Net}}, b_{\text{seg}}, \mathscr{L}_{\text{U-Net}})$ {train on n_{seg} batches from b_{seg}, backpropagating $\mathscr{L}_{\text{U-Net}}$}
  $\texttt{update}(\mathscr{L}_{\text{U-Net}})$
  $\mathscr{L}_{\text{UOLO}} \leftarrow \mathscr{L}_{\text{YOLO}} + \mathscr{L}_{\text{U-Net}}$

of the predicted bounding box $\hat{b}$,

$$\begin{aligned}(\hat{b}_x, \hat{b}_y) &= (\sigma(\hat{x}) + x_{i,j,k}, \sigma(\hat{y}) + y_{i,j,k}) \\ (\hat{b}_w, \hat{b}_h) &= (w_{i,j,k} e^{\hat{w}}, h_{i,j,k} e^{\hat{h}}) \\ \text{confidence} &= \sigma(\hat{c})\end{aligned} \tag{3.2}$$

where σ is a sigmoid function. YOLOv2 is trained by optimizing the loss function:

$$\mathscr{L}_{\text{YOLO}} = \lambda_1 \mathscr{L}_{\text{centers}} + \lambda_2 \mathscr{L}_{\text{dimensions}} + \lambda_3 \mathscr{L}_{\text{confidence}} + \lambda_4 \mathscr{L}_{\text{classes}} \tag{3.3}$$

where λ_i are predefined weighting factors, $\mathscr{L}_{\text{centers}}$, $\mathscr{L}_{\text{dimensions}}$ and $\mathscr{L}_{\text{confidence}}$ are mean squared errors, and $\mathscr{L}_{\text{classes}}$ is the cross-entropy loss. Each loss term penalizes a different error: 1) $\mathscr{L}_{\text{centers}}$ penalizes the error in the center position of the cells; 2) $\mathscr{L}_{\text{dimensions}}$ penalizes the incorrect size, *i.e.* height and width, of the bounding box; 3) $\mathscr{L}_{\text{confidence}}$ penalizes the incorrect prediction of a box presence; 4) $\mathscr{L}_{\text{classes}}$ penalizes the misclassification of the objects.

3.2.3 UOLO for joint object detection and segmentation

Let $M_{\text{U-Net}}$ be an U-Net-like network that, given pairs of images and binary masks, can be trained for performing segmentation by minimizing $\mathscr{L}_{\text{U-Net}}$ (Eq. 3.1). $M_{\text{U-Net}}$ has a second output corresponding to the concatenation of the downsampled decoding maps with its bottle neck (last encoder layer). The resulting tensor corresponds to a set of multi-scale representations of the original image that are supplied to the object detection block D_{YOLO}, which, by its turn, can be optimized via $\mathscr{L}_{\text{YOLO}}$, defined in Eq. 3.3. D_{YOLO} and $M_{\text{U-Net}}$ are then merged by concatenation into M_{UOLO}, a single model that can be optimized by minimizing the addition of the corresponding loss functions:

$$\mathscr{L}_{\text{UOLO}} = \mathscr{L}_{\text{YOLO}} + \mathscr{L}_{\text{U-Net}} \tag{3.4}$$

Thanks to the straightforward definition of the loss function in Eq. (3.4), M_{UOLO} can be trained with a simple iterative scheme detailed in Algorithm 1. In essence, $\mathscr{L}_{\text{U-Net}}$ is updated only when segmentation information is available. However, a global weight update is performed at every step based on the prediction error backpropagation. Furthermore, the outlined training scheme allows for a different number of strong (pixel-wise) and weak (bounding boxes) annotations, easing its application to medical images.

3.3 Experiments and results

3.3.1 Datasets and experimental details

All images are cropped around the field-of-view (FOV) (determined via Otsu's thresholding) and resized to 256×256 pixels. The side of the square ground truth bounding boxes is set to 32 and 64 for the FV and OD, respectively, following their relative size in the image. For training, n_{det} and n_{seg} (Algorithm 1) are set to 256 and 32, respectively. Online data augmentation, a mini-batch size of 8, and the Adam optimizer (learning rate of 1×10^{-4}) were used for training, while 25% of the data was kept for validation. The bounding box with highest confidence for each class is kept. The predicted soft segmentations are binarized using a threshold of 0.5.

We test UOLO on 3 public eye fundus datasets with healthy and pathological images: 1) Messidor (Decenciere *et al.*, 2014) has 1 200 images (1440×960, 2240×1488 and 2304×1536 pixels, 45° FOV), 1 136 having ground truth for OD segmentation and FV centers*; 2) IDRID† training set has 413 images (4288×2848 pixels, 50° FOV) with OD and FV centers and 54 with OD segmentation; 3) DRIVE (Staal *et al.*, 2004) has 40 images (768×584 pixels, 45° FOV) with OD segmentation‡.

The OD segmentation is evaluated with intersection over union (IoU) and Sorensen-Dice coefficient overlap metrics. The detection is evaluated in terms of mean euclidean distance (ED) between the prediction and the ground truth. We also evaluate ED relatively to the OD radius (R), $\bar{D} = ED/R$ (Gegundez-Arias *et al.*, 2013; Aquino, 2014). As the OD may vary significantly in different images of the same size (from 0.1 to 0.2 of the retinal size) (Li and Chutatape, 2001), R is estimated considering the middle value (0.15) as follows:

$$R = 0.15 \, \frac{D_{FOV}}{2} \tag{3.5}$$

where D_{FOV} is the width of the FOV, serving as an approximation of the retinal diameter (Gegundez-Arias *et al.*, 2013). Finally, detection success, S_{1R}, is assessed using the maximum distance criteria of 1 OD radius, *i.e.*, a detection is considered successful if the detected center is at a distance from the ground truth center of less than 1 OD radius, and unsuccessful otherwise.

*http://www.uhu.es/retinopathy

†https://idrid.grand-challenge.org/, available since January 20, 2018

‡https://sites.google.com/a/uw.edu/src/useful-links

Table 3.1: UOLO performance on optic disc (OD) detection and segmentation and fovea (FV) detection. *n*: number of training images for detection and segmentation.

Datasets		**n**		**OD seg.**		**OD det.**		**FV det.**	
Train	**Test**	seg.	det.	IoU	Dice	$\bar{D}$	S_{1R}	$\bar{D}$	S_{1R}
Messidor		680	680	0.88	0.93	0.111	99.74	0.121	99.38
Messidor		100	680	0.87	0.93	0.114	99.74	0.114	97.89
IDRID		30	280	0.88	0.93	0.095	99.79	0.288	93.78
Messidor	IDRID	852	852	0.84	0.91	0.138	99.78	0.403	89.06
Messidor	DRIVE	852	852	0.82	0.89	0.171	97.50	-	-

3.3.2 Results and discussion

We evaluate UOLO both inter and intra-dataset-wise. For inter-dataset experiments, UOLO was trained on Messidor and tested in the other datasets whereas for intra-dataset studies stratified 5-fold crossvalidation was used. Using this crossvalidation scheme allows to ensure that each sample appears exactly once in testing, and that each fold contains approximately the same percentage of samples of each class as the complete set. We do not extensively optimize the training parameters so we can verify how robust UOLO is when dealing with segmentation and detection simultaneously.

Table 3.1 shows the results of UOLO for the OD detection and segmentation and FV detection tasks, as well as the results from the experiment to assess the influence of the amount of annotations for OD segmentation in the training data, since these are the more costly to obtain. Instead of using the available 680 Messidor images with OD segmentation ground truth, only 100 of these were used for training. Table 3.2 compares our performance with state-of-the-art methods and Fig. 3.3 shows two prediction examples in complex detection and segmentation cases. Fig. 3.4 presents more examples of UOLO predictions. Note that for intra-dataset experiments the values shown in the tables are the average of the 5 folds' results.

UOLO achieves equal or better performance in comparison to the state-of-the-art on both

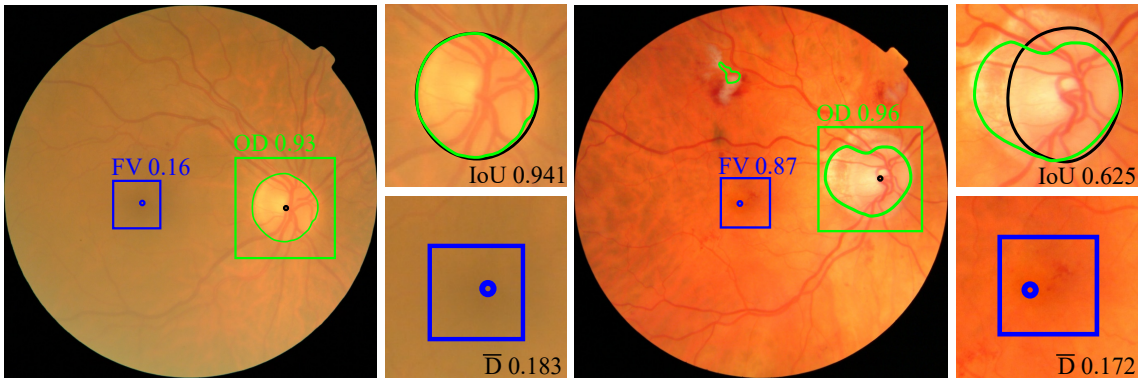
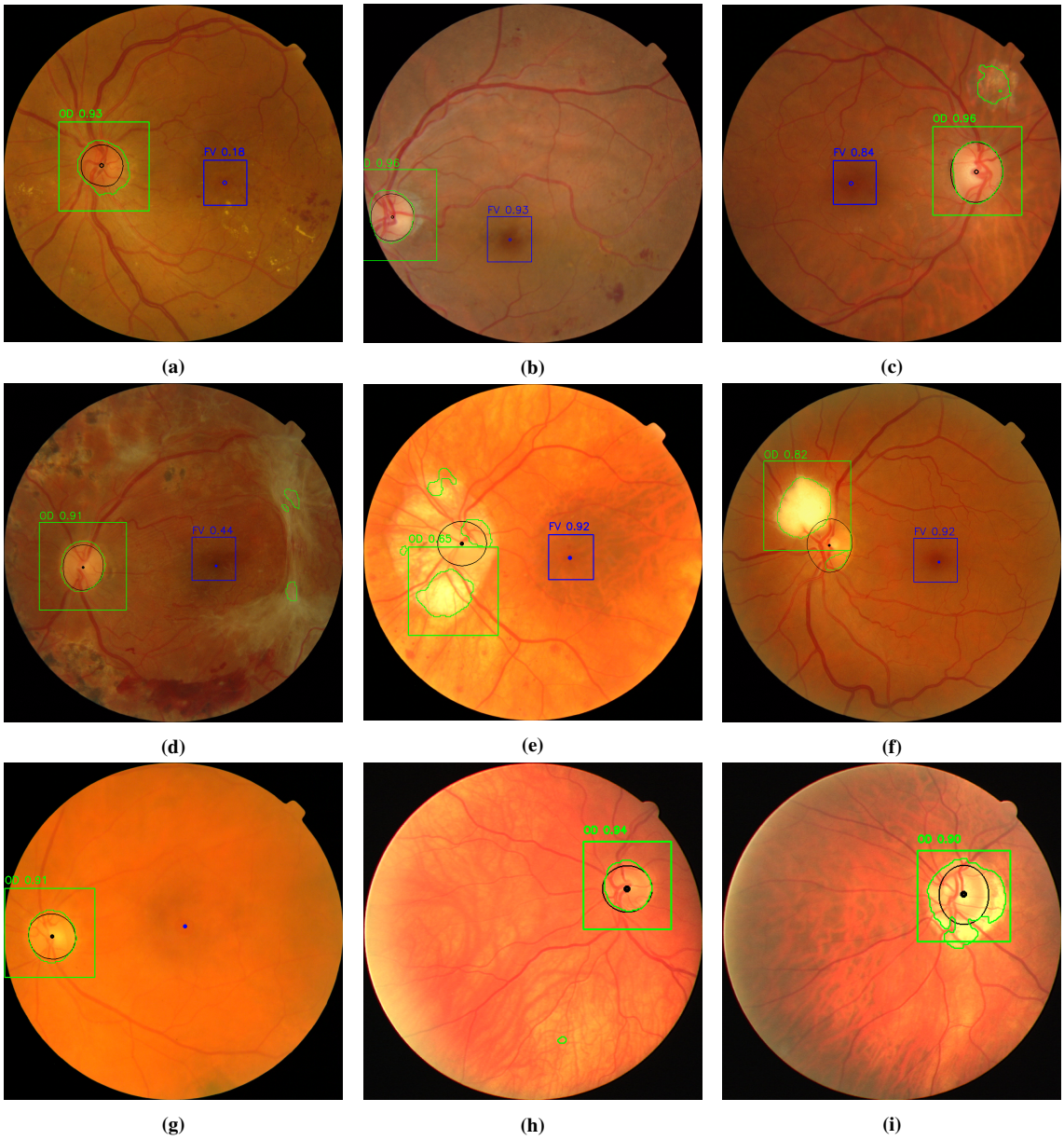


Figure 3.3: Examples of results of UOLO on Messidor images. Green curve: segmented optic disc (OD), green and blue boxes: predicted OD and FV locations, respectively; black curve: ground truth OD segmentation; black and blue dots: ground truth OD and FV locations, respectively. The object detection confidence is shown next to each box. IoU and normalized distance ($\bar{D}$) values are also shown.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3.4: Examples of results of UOLO on Messidor and DRIVE dataset images. Green curve: segmented optic disc (OD), green and blue boxes: predicted OD and FV locations, respectively; black curve: ground truth OD segmentation; black and blue dots: ground truth OD and FV locations, respectively. The object detection confidence is shown next to each box. a-f: Messidor; h and i: DRIVE dataset.

Table 3.2: State-of-the-art for OD detection and segmentation and FV detection.

(a) OD segmentation

Dataset	Messidor		DRIVE	
Method	IoU	Dice	IoU	Dice
UOLO	0.88	0.93	0.82	0.89
U-Net	0.88	0.93	0.81	0.88
Dai *et al.* (2017)	0.91	-	-	-
Dashtbozorg *et al.* (2015)	0.89	0.94	-	-
Roychowdhury *et al.* (2016)	0.84	-	0.81	-
Morales *et al.* (2013)	0.82	-	0.72	-
Salazar-Gonzalez *et al.* (2014)	-	-	0.82	-
Yu *et al.* (2012)	0.83	-	-	-
Aquino *et al.* (2010)	0.86	-	-	-

(b) OD and FV detection

Task	**OD detection**				**FV detection**	
Dataset	Messidor		DRIVE		Messidor	
Method	ED	S_{1R}	ED	S_{1R}	ED	S_{1R}
UOLO	9.40	99.74	8.13	97.5	10.44	99.38
YOLOv2	6.86	100	7.20	97.5	9.01	100
Al-Bander *et al.* (2018)	-	97	-	-	-	96.6
Aquino (2014)	-	-	-	-	16.09	98.24
Gegundez-Arias *et al.* (2013)	-	-	-	-	20.17	98.24
Aquino *et al.* (2010)	-	98.83	-	-	-	-
Kamble *et al.* (2017)	23.17	99.75	15.57	100	34.88	99.40
Dashtbozorg *et al.* (2015)		99.75	-	-	-	-
Yu *et al.* (2012)	-	99.10	-	-	-	-

detection and segmentation tasks (IoU 0.88 ± 0.09 on Messidor) in a single step prediction. Furthermore, the proposed network is robust even in inter-dataset scenarios, maintaining both segmentation and detection performances. This indicates that the abstract representations learned by UOLO are highly effective for solving the detection and segmentation tasks. Indeed, in the OD detection task, the study of Kamble *et al.* (2017), which reports performance on both datasets (Table 3.2), shows an euclidean distance between predicted and ground truth center points of 23.17 and 15.57 pixels in Messidor and DRIVE, respectively, comparing with UOLO's 9.40 and 8.13 pixels. Our method shows, thus, both a very good performance and high consistency across datasets.

UOLO presents a success rate (S_{1R}) for OD detection higher than 97.5% across all datasets, and above 89% for FV detection (Table 3.1). On Messidor, the UOLO's success rates for OD and FV detection are 99.74 and 99.38 %, respectively, surpassing the results reported by Al-Bander *et al.* (2018), whose approach is also deep learning-based, and presented S_{1R} of 97% and 96.6%, respectively.

Also, it is worth noting that our segmentation and detection performances do not alter significantly even when UOLO is trained with only 15% of the pixel-wise annotated images (Table 3.1).

This means that UOLO does not require a significant amount of pixel-wise annotations, easing its application on the medical field, where these are particularly expensive to obtain.

Our results also suggest that UOLO is capable of using multi-scale information (*eg.* relative position to the OD or vessel tree) to perform predictions. For instance, Fig. 3.3 shows UOLO's output for two Messidor images, illustrating that the network is capable of detecting the FV in a low contrast scenario. On the other hand, the segmentation and detection processes are not completely interdependent, as expected from the proposed training scheme, since the network segments OD confounders outside the detected OD region (right image from Fig. 3.3 and Fig. 3.4c-e). Another advantage of UOLO is that these segmentation errors could be corrected by limiting the pixel-wise predictions to the found OD region. Unlike hand-crafted feature-based methods, UOLO does not require an extensive parameter tuning, since this was not performed and the results were very promising, and it is simple to extend to different applications.

There are cases, however, in which regions that are close to the OD but are brighter than the OD are incorrectly detected and segmented as such (Fig. 3.4e and 3.4f). The locations of these false OD regions do not fall far from the expected given the FV locations, and given the color similarity it can be understood why the algorithms mispredicts in such cases. Additionally, in cases such as the one from Fig. 3.4g, in which the region surrounding the FV appears darker than the FV itself, UOLO fails to detect the FV. However, the fact that the algorithm does not mandatorily detects these structures when they are not visible in the image, such as in Fig. 3.4h and 3.4i, is a positive feature.

We also evaluate U-Net ($\mathscr{M}_{\text{U-Net}}$, Fig. 3.2) for OD segmentation and YOLOv2 (with a pre-trained Inception-v3 as feature extractor) for OD and FV detection (Table 3.2). The training conditions were set as in UOLO. UOLO segmentation performance is practically the same as U-Net, whereas the detection drops slightly when comparing with YOLOv2, mainly for OD detection. However, one has to consider the trade-off between computational burden and performance, since UOLO network has 23 347 063 parameters, whereas U-Net has 15 063 985 and YOLOv2 has 21 831 470, being that for training U-Net and YOLO a total of 36 895 455 parameters have to be optimized (60% increase).

3.4 Concluding remarks

In this chapter we presented UOLO, a novel network that performs joint detection and segmentation of objects of interest in medical images by using the abstract representations learned by U-Net. Furthermore, UOLO can detect objects from a different class for which segmentation ground truth is available. We tested UOLO for simultaneous fovea detection and optic disc detection and segmentation, achieving state-of-the-art results. This network can be trained with relatively few images with segmentation ground truth and still maintain a high performance. UOLO is also robust to inter-dataset settings, thus showing great potential for applications in the medical image analysis field.

Chapter 4

DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images

This chapter in based on the following publication:
Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., *et al.* DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715, 7 2020c. doi: 10.1016/j.media.2020.101715

Diabetic retinopathy (DR) grading is crucial in determining the adequate treatment and follow up of patients, but the screening process can be tiresome and prone to errors. Deep learning approaches have shown promising performance as computer-aided diagnosis (CAD) systems, but their black-box behaviour hinders clinical application. In this chapter, we propose DR|GRADUATE, a novel deep learning-based DR grading CAD system that supports its decision by providing a medically interpretable explanation and an estimation of how uncertain that prediction is, allowing the ophthalmologist to measure how much that decision should be trusted. We designed DR|GRADUATE taking into account the ordinal nature of the DR grading problem. A novel Gaussian-sampling approach built upon a Multiple Instance Learning framework allows DR|GRADUATE to infer an image grade associated with an explanation map and a prediction uncertainty while being trained only with image-wise labels. DR|GRADUATE was trained on the Kaggle DR detection training set and evaluated across multiple datasets. In DR grading, a quadratic-weighted Cohen's kappa (κ) between 0.71 and 0.84 was achieved in five different datasets. We show that high κ values occur for images with low prediction uncertainty, thus indicating that this uncertainty is a valid measure of the predictions' quality. Further, bad quality images are generally associated with higher uncertainties, showing that images not suitable for diagnosis indeed lead to less

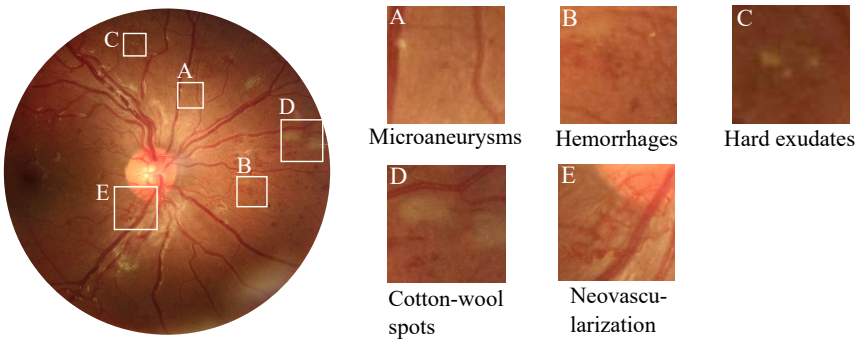


Figure 4.1: Examples of typical retinal lesions characteristic of diabetic retinopathy.

trustworthy predictions. Additionally, tests on unfamiliar medical image data types suggest that DR|GRADUATE allows outlier detection. The attention maps tend to highlight regions of interest for diagnosis. These results show the great potential of DR|GRADUATE as a second-opinion system in DR severity grading.

4.1 Introduction

DR is generally classified according to the absence or presence of new abnormal vessels as non-proliferative DR (NPDR) or proliferative DR (PDR), respectively. NPDR can be further graded as mild, moderate and severe (The Royal College of Ophthalmologists, 2012). These stages can be ordered according to their risk of progression (Arévalo *et al.*, 2013). Namely, Table 4.1 represents the International Clinical DR Scale (Wilkinson *et al.*, 2003) (from Table 2.3), herein adapted to consider only one image for diagnosis (Porwal *et al.*, 2020) by removing the quadrant concept, since in the case of a single fundus image analysis quadrant assessment is limited due to the under-representation of some of the quadrants. Thus, in practice, the exact counting of lesions in each quadrant is not performed by the specialists, but rather the estimation of what happens in the full eye based on their experience and medical knowledge.

Mild NPDR is characterized by the presence of microaneurysms (MAs), which are the first sign of DR. Hemorrhages (HEMs) start to appear with the evolution of the disease in the deeper layers of the retina. The disease progression results in an increase of MAs, dot-hemorrhages (red lesions), retinal thickening, hard exudates (EXs), and the appearance of intraretinal microvascular abnormalities (IrMAs), venous beading and cotton-wool spots (CWSs), also known as soft exudates, in the innermost retina. Finally, PDR results from abnormal vessel formation, *i.e.*, neovessels (NVs). Other signs of PDR are the vitreous/pre-retinal hemorrhages (PHEMs) and pre-retinal fibrosis (PFIB) (Hall, 2011). Fig. 4.1 shows examples of some of these lesions.

Computer-aided diagnosis (CAD) systems can improve the DR screening pipeline both reducing the burden (Scotland *et al.*, 2007) and by proving a second objective opinion to the ophthalmologists, reducing subjectivity in the diagnosis (Gargeya and Leng, 2017; Abràmoff *et al.*, 2016; Gulshan *et al.*, 2016; Quellec *et al.*, 2017). Namely, deep learning has recently allowed for

Table 4.1: Diabetic retinopathy (DR) severity scale. NPDR: Non-proliferative DR; PDR: proliferative DR. Non-referable DR: R0 and R1; Referable DR: R2, R3 and R4.

Grade		Description
R0 -	No DR	No visible sign of abnormalities
R1 -	Mild NPDR	Presence of microaneurysms only
R2 -	Moderate NPDR	More than just microaneurysms but less than severe NPDR
R3 -	Severe NPDR	Any of the following: >20 intraretinal hemorrhages Venous beading Intraretinal microvascular abnormalities no signs of PDR
R4 -	PDR	Either or both of the following: Neovascularization Vitreous/pre-retinal hemorrhage

these systems to achieve near-human performance in DR detection, *i.e.* detection of at least one DR-related lesion. On the other hand, DR grading, *i.e.* staging of the pathology according to its severity, is a more complex problem since it requires the identification and integration of different abnormalities. In fact, specialists tend to disagree on the grade of complex cases (Krause *et al.*, 2018) and thus the external opinion of CAD systems may contribute to further improve the success of DR grading. However, due to this higher complexity and the inherent black-box behaviour of deep learning systems, it may be difficult for specialists to understand and accept the system's decision. With this in mind, we propose a novel DR grading CAD system that not only supports its decision by providing a medically interpretable explanation but also estimates how uncertain that prediction is. By doing so, the ophthalmologist not only has a reasoning behind the system's decision, but also a measure of how much that decision should be trusted, making the DR diagnosis process less prone to errors.

4.1.1 State-of-the-art

Deep learning has become the preferred approach, over field knowledge-based feature engineering approaches, for DR grading since it allows to better exploit the large number of data available and to better deal with the labeling noise resulting from the complexity of the task. For instance, Krause *et al.* (2018) fine-tuned an Inception-v4 with 1 600 000 images of size 779×779 pixels to avoid the elimination of small lesions such as MAs. Also, to account for data imbalance, the authors apply a cascade of thresholds on the output probabilities of the model for each DR grade to obtain the final image classification. These thresholds are sequentially applied in a severity descending order to ensure that a minimum grade-wise sensitivity is achieved. Instead, de la Torre *et al.* (2018) considered DR grading as an ordinal classification problem and thus aimed at reducing the distance between predicted and true labels of an image. For that, they proposed a quadratic-weighted-Kappa-based loss function, which allowed to improve the performance of a DR grading network comparing to training with the cross-entropy loss. The authors tested different

input images sizes (squares with side of 128, 256, 384 and 512 pixels), and the highest resolution yielded the best performance. The network, which has 11.3 million parameters for the 512×512 case, was trained in a public dataset (Kaggle DR detection). To overcome the high imbalance of the dataset, the data is artificially balanced using data augmentation, oversampling the least represented classes.

Despite the high classification performance of these systems, their black-box behavior hinders the application on screening settings. Indeed, providing explanations for DR grading is challenging due to the complexity of the task, and thus research is mainly focused on the binary detection and referral tasks. For instance, Quellec *et al.* (2017) analyzed several explanation methods for the referable DR decisions of their CNNs'. Namely, the authors compare deconvolution (Zeiler and Fergus, 2014), sensitivity analysis (Simonyan *et al.*, 2014) and layer-wise relevance propagation (Bach *et al.*, 2015). Based on (Simonyan *et al.*, 2014), the authors also proposed the creation of a pixel-wise explanation map that is jointly optimized with the CNN prediction during training. Costa *et al.* (2019) proposed the attention map generation via a pre-trained network coupled with a Multiple Instance Learning (MIL)-based approach for DR detection. Namely, the network is cropped at early layers to reduce the model's receptive field, and the classification is performed by a 1×1 convolution and global max-pooling, following the assumption that the presence of one unhealthy patch is enough to label the entire image. Because of this, the output of this convolution is a map where each element indicates the malignancy score of a corresponding patch of the input image, thus explaining why the image label was predicted. Instead, González-Gonzalo *et al.* (2018) has applied an iterative inpainting approach that, at each step, allows to identify progressively less discriminative pixels of the image for DR classification. This method involves the computation of the saliency map at each step, *i.e.*, the derivative of the classification output with respect to the input image using guided back-propagation. The authors use a pre-trained VGG-16 on the Kaggle dataset (512×512 pixel images). Similarly to other grading approaches, the problem is treated as being ordinal and thus the model is optimized using a mean-square error loss. To avoid overfitting due to the high class imbalance, the classes are balanced taking into account the referal/non-referal sample distribution.

Besides explaining the decision, uncertainty estimation of a model's prediction is of interest as it helps to avoid consequences of the blind use of the model response (Kendall and Gal, 2017). This is particularly true on medical settings, where misdiagnosis can have severe consequences for the patients. With this in mind, Leibig *et al.* (2017) analysed the uncertainty information from deep neural nets for DR detection and referable DR detection. The authors tested the dropout-based Bayesian uncertainty estimation, comparing it with alternative techniques, such as directly analysing the network's softmax output. In the dropout-based approach, dropout is switched on during inference. Performing multiple predictions on the same input image can be interpreted as having an ensemble of highly similar models, and thus disagreement between the inferences is indicative of the global model uncertainty (Gal and Ghahramani, 2016). The authors state that the Bayesian treatment worked better than the softmax output as a proxy for uncertainty and also show that uncertainty-aware decision referral can improve the diagnosis.

4.1.2 Contributions

Several research works have already addressed DR grading, explainability and uncertainty estimation separately. Instead, in this study we propose DR|GRADUATE, a system that is capable of simultaneously providing an explanation and a measure of uncertainty together with a DR severity grade label. Our main contributions are the following:

- estimation of a DR grade with an associated uncertainty, thus allowing to better establish which cases require further inspection by specialists, which is of high interest in computer-aided diagnosis systems;
- explanation of the algorithm decision, easily and directly interpretable by the retinal specialists, in the form of an attention map, which will promote its use by medical professionals since it partially mitigates the black-box behaviour associated with deep nets;
- extensively validated deep learning-based method for DR severity grading of eye fundus images in the internationally recognized DR levels.

4.2 DR|GRADUATE for DR grading

The backbone of DR|GRADUATE, depicted in Fig. 4.2, is composed of several convolutional-batch normalization blocks interleaved with max-pooling layers that have already shown potential for DR grading tasks (de la Torre *et al.*, 2017). The top of the network is composed of a novel component that allows for the inference of the image's grade $\hat{y}_g$ together with the corresponding explainability $\mathscr{E}$ and the explicit uncertainty u of the decision, without using labels other than the DR grade, such as lesion-wise annotations.

Let DR|GRADUATE be a classification model that predicts $\hat{y}_g = argmax(p)$, where p_c is the class-wise probability and $\sum_c^C p_c = 1$, with C denoting the number of considered classes. The following assumptions regarding DR grading are made: [`assumption 1`] grading is a discrete ordinal classification problem with an image label $y \in \mathscr{G} = \{0,1,2,3,4\}$, where $\mathscr{G}$ is the set of possible DR grades; in this scenario, we want to optimize DR|GRADUATE by minimizing the distance between the target and the prediction, $min(|y - \hat{y}_g|)$, and [`assumption 2`] there is always a different lesion type per grade on the universe of possible DR-related lesions, $DR_{lesions}$; thus, if one associates a numerical value to each lesion from $DR_{lesions}$, accordingly to the DR stage in which they appear, $\hat{y}_g$ can be predicted by following the standard multiple instance learning (MIL) assumption, *i.e.* $max(DR_{lesions}) \mapsto \hat{y}_g$.

DR grading can be seen as a regression problem and consequently the predicted image grade $\hat{y}_g$ can be inferred from a continuous output $\hat{y}_r \in \mathbb{R}_+$. This also allows for the implicit MIL of the problem, as we can retrieve $\hat{y}_r$ from a predicted lesion map L, $\hat{y}_r = max(L)$. Furthermore,

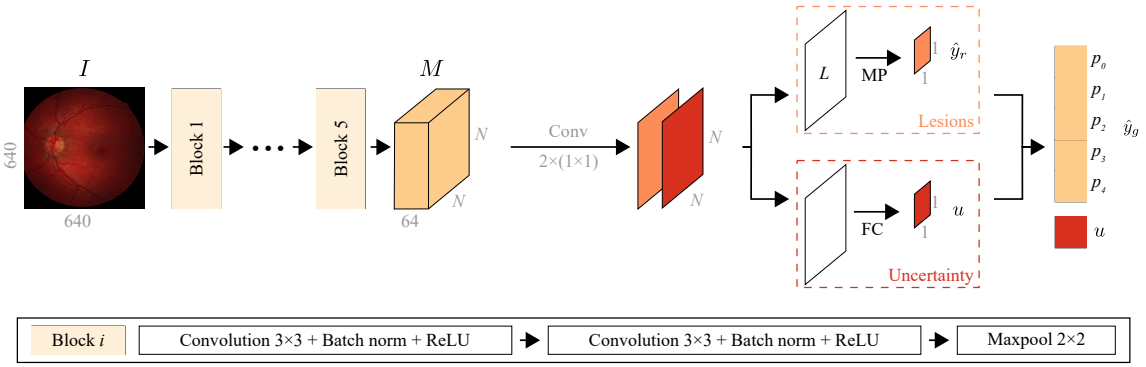


Figure 4.2: DR|GRADUATE's architecture. The network is composed of several convolutional-batch normalization blocks interleaved with max-pooling layers. For each input image I, DR|GRADUATE predicts a diabetic retinopathy (DR) grade $\hat{y}_g$ and a grade uncertainty u. M is the output of the last layer of DR|GRADUATE's backbone after assessing I. The lesion map L is computed by performing a $1 \times 1 \times 1$ convolution with a linear activation over M. L is a map where the value of each element L_{ij} indicates the presence of a lesion of grade $\lfloor L_{ij} \rceil^{\dagger}$ on a patch of size dependent on the model's receptive field in the input. Following the Multiple Instance Learning approach, the grade of an image is initially estimated as $\hat{y}_r = max(L)$. The output probabilities ($p_c, c \in \{0,...,4\}$) are then computed from $\hat{y}_r$ based on a Gaussian distribution centered on $\hat{y}_r$ with variance equal to the uncertainty u, with u learned during training. (MP: Max-pooling; FC: fully connected).

the uncertainty u could be estimated by $u \propto |y - \lfloor \hat{y}_r \rceil|^{\dagger}$. However, this uncertainty does not properly hold in scenarios where the regression cut-off thresholds are adjusted to better fit the data distribution and malignancy priority, as in many Kaggle competition solutions[1]. Also, in terms of clinical applicability, it is of higher interest to provide a class-wise probability to clinicians since it is easier to interpret. Because of this, [`assumption 3`] DR|GRADUATE assumes that the prediction and its uncertainty can be modeled by a Gaussian curve centered near the most probable class. Namely, for each image, DR|GRADUATE predicts a variance σ^2, which allows for $\hat{y}_r \in \mathbb{R}_+ \mapsto \hat{y}_g = argmax(p)$. With this, DR|GRADUATE is capable of providing an uncertainty σ^2 and a class-wise probability that still attempts at minimizing the distance between the prediction and the ground truth. The next sections describe DR|GRADUATE with more detail.

4.2.1 Predicting a grade with uncertainty

For each input image I, DR|GRADUATE predicts a DR grade $\hat{y}_g$ and a grade uncertainty u. Let M be the output of the last layer of DR|GRADUATE's backbone after assessing an input image I. M has size $N \times N \times F$, where N and F are, respectively, the side and number of features. The lesion map L is computed by performing a $1 \times 1 \times 1$ convolution with a linear activation over M, so that $L \in \mathbb{R}$ has size $N \times N \times 1$. L is a map where the value of each element L_{ij} indicates the presence of a lesion of grade $\lfloor L_{ij} \rceil^{\dagger}$ on a patch of size dependent on the model's receptive field in the input. Following the MIL approach, the grade of an image is initially estimated as $\hat{y}_r = max(L)$ (Costa *et al.*, 2019).

[†] $\lfloor \rceil$ denotes rounding to the nearest integer, $\lfloor \rfloor$ rounding down and $\lceil \rceil$ rounding up.

[1] `https://storage.googleapis.com/kaggle-forum-message-attachments/88655/2795/competitionreport.pdf`

Having into account `assumptions 1` and 3, $\hat{y}_r \mapsto \hat{y}_g$ must ensure that $p_{\lfloor \hat{y}_r \rfloor} > p_{\lfloor \hat{y}_r \rfloor - 1} > p_{\lfloor \hat{y}_r \rfloor - 2} > \dots$, and $p_{\lceil \hat{y}_r \rceil} > p_{\lceil \hat{y}_r \rceil + 1} > p_{\lceil \hat{y}_r \rceil + 2} > \dots$, *i.e.*, that the classes near to $\hat{y}_r$ have higher probability than the others. For each image, DR|GRADUATE must thus predict a probability distribution biased around the classes closer to $\hat{y}_r$. This bias is introduced in the model by means of a Gaussian distribution $\mathcal{N}$ centered on $\hat{y}_r$ and thus the image-wise grade probabilities p are:

$$
\begin{aligned}
p' &= \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-0.5 \frac{[i - \hat{y}_r]^2}{\sigma^2} \right), \forall c \in \mathcal{G} \right\} \\
p &= \frac{p'}{\sum_{c \in \mathcal{G}} p'_c}
\end{aligned}
\tag{4.1}
$$

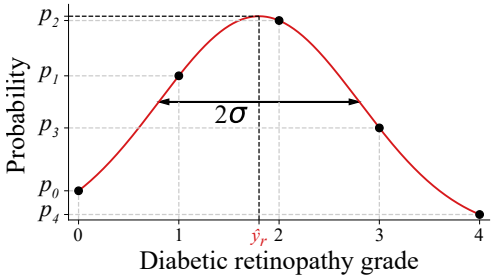
where σ^2 is the variance of $\mathcal{N}$, $p \in [0,1]$, and p' is a set of the non-normalized grade probabilities, p'_c. Fig. 4.3a illustrates the sampling of the probability values p from the Gaussian probability distribution. Knowing that the uncertainty of the prediction can be measured by the value of its entropy (Friston, 2010), *i.e.* u is directly related to $S = -\sum_c p_c log(p_c)$, it follows that since S depends on σ^2, then u depends on σ^2. Thus by estimating σ^2 it is possible to infer the uncertainty associated with each predicted grade, $\hat{y}_g = argmax(p)$.

In DR|GRADUATE, σ^2 is inferred by using a fully-connected layer on top of M. Since both $\hat{y}_r$ and σ^2 depend on M, both grading and the associated uncertainty affect the backbone of the model, easing its convergence. Indeed, DR|GRADUATE is trained end-to-end minimizing the loss function:

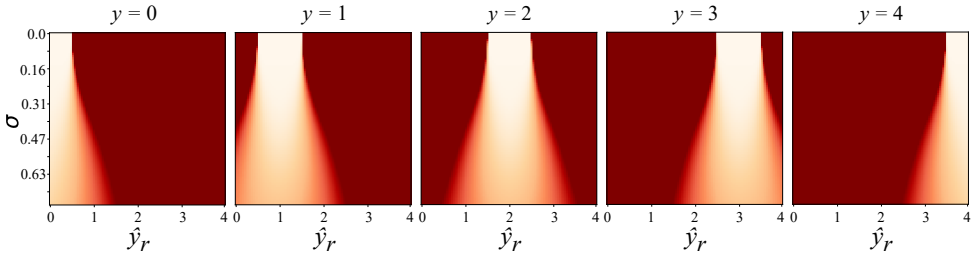
$$
\mathcal{L} = -\alpha \sum_{c \in \mathcal{G}} [b_c log(p_c)] + (1 - \alpha)\sigma^2 \tag{4.2}
$$

where b_c is the hot-encoded (*i.e.*, a binary representation of) i and $\alpha \in [0,1]$ is a weighting factor. The first term, the categorical cross-entropy, leads DR|GRADUATE to predict $\hat{y}_r$ and thus $\hat{y}_g$ as close to the truth as possible. The second term regularizes σ^2 to avoid excessive Gaussian spreads and thus the stagnation of the learning on the local minimum of near equiprobable predictions.

The global loss $\mathcal{L}$ and the structure of DR|GRADUATE implicitly introduce an inter-dependency between $\hat{y}_r$ and σ that allows to deal with images of different complexity. For less complex images, both σ and the preliminary class prediction error, $|y - \lfloor \hat{y}_r \rceil|$, should be low. However, for difficult images, high values of σ^2 can compensate for classification errors since the higher spread of the Gaussian tends to increase the predicted probability of the correct class, lowering the cross-entropy loss. Fig. 4.3b shows the evolution of $\mathcal{L}$ (Eq. 4.2) as function of $\hat{y}_r$ and σ, for different targets y, $y \in \mathcal{G}$. These plots show that when $\hat{y}_r$ is close to y the $\mathcal{L}$ loss reaches its minimum. Further, if one freezes $\hat{y}_r$ at a value close to the ground truth, the σ that leads to a lower $\mathcal{L}$ loss is smaller than if $\hat{y}_r$ is fixed at a value far from the ground truth.



(a) Probability Gaussian distribution with mean $\hat{y}_r$ and standard deviation σ (Eq. 4.1).



(b) $\mathscr{L}$ loss values (Eq. 4.2, with $\alpha = 0.7$) as function of $\hat{y}_r$ and σ, for $y \in \mathscr{G}$. Loss colorbar: 0 2 .

Figure 4.3: Uncertainty estimation (Eq. 4.1) and loss representation (Eq. 4.2).

4.2.2 Explainability of the DR grade

The goal of this study does not include a pixel-level prediction of lesions. Even human-performed lesion segmentation shows high variability, due to the extremely small size of some lesions and also to differences in the annotation style. Indeed, some ophthalmologists prefer to only annotate clusters of lesions, whereas others are more detailed and prefer to give lesion-wise markings. Since DR image grading is performed based on the presence of the typical DR lesions, pinpointing relevant lesions should be a valuable tool for doctors to assess the algorithm's prediction without the need for a refined segmentation.

DR|GRADUATE provides a grade-wise explanation map, as illustrated in Fig. 4.4. For each grade c the lesion presence encoded in L is translated to an explainability map $\mathscr{E}_c$ of the same size of I via:

$$\mathscr{E}'_c(s\cdot i, s\cdot j) = \begin{cases} 1 & \text{if } c-0.5 \leq L_{ij} < c+0.5 \\ 0 & \text{otherwise} \end{cases}$$
$$\mathscr{E}_c = \mathscr{E}'_c \circledast \mathscr{N}_{2,\sigma'} \tag{4.3}$$

where $\mathscr{E}'_c$ is a matrix of the same size of $\mathscr{E}_c$, $c \in \mathscr{G} \setminus \{0\}$ is one of the possible pathological grades, i, j are each row and column position, respectively, in the L map of side N ($i, j \in \mathbb{N} \mid i, j < N$), and $\mathscr{N}_{2,\sigma'}$ is a convolutional kernel from a two-dimensional multivariate normal distribution with standard deviation σ', with σ' related to the receptive field of the network, and s is the stride of DR|GRADUATE. The convolution with the Gaussian kernel allows to give higher priority to

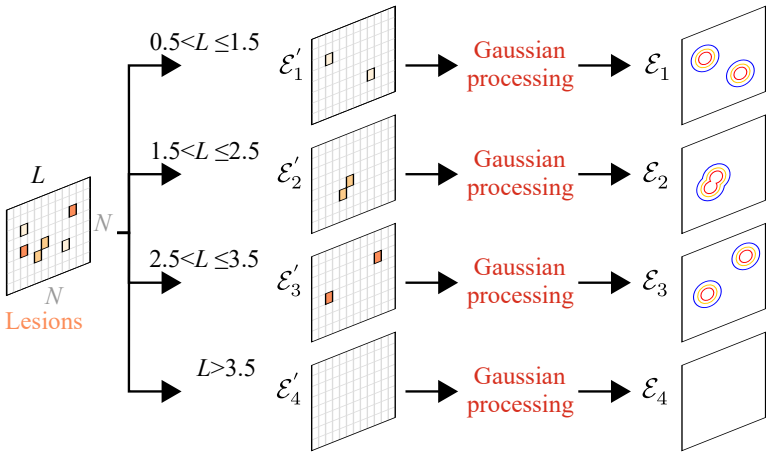


Figure 4.4: Attention maps construction. Gaussian processing is described in Eq. 4.3.

central pixels of the receptive field, which have a higher contribution to the value of each L_{ij}.

4.2.3 Dealing with class imbalance

DR grading is an imbalanced problem skewed towards the healthy cases, similarly to many other medical imaging problems. Data imbalance introduces convergence issues during model training, demanding for class balancing strategies. Oversampling has shown to be an overall good choice for dealing with class imbalance (Buda *et al.*, 2018). However, excessively reducing class imbalance may lead to an unnecessarily high trade-off between the model's sensibility and specificity, *i.e.* increase the misclassifications of healthy images, since the most represented classes will now tend to be relatively over-classified. With this in mind, DR|GRADUATE is trained using an iterative batch balancing scheme that has already shown success for DR grading[2]:

$$w_{ct} = r^{t-1} w_{co} + (1 - r^{t-1}) w_{cf} \tag{4.4}$$

where w_{ct} is the expected ratio of images of grade c on the batch at epoch t, w_{c0} is the representativeness of class c on the dataset, w_{cf} is the ratio of class c on the batch after f training epochs and r is the decay rate that controls the update of w_{ct}. With this, in the beginning of the training the network is confronted with a completely balanced set and does the initial learning based on this, and then progressively starts to adjust (or fine-tune) to a more real data distribution, but still not as skewed as the original training set distribution. The oversampling of the least represented classes is performed via online horizontal and vertical flips, rotations, brightness adjustments and contrast normalization.

[2] https://www.kaggle.com/blobs/download/forum-message-attachment-files/2797/report.pdf

4.3 Experiments

Intra- and inter-datasets experiments were performed for evaluating DR|GRADUATE in terms of: 1) DR grading performance, 2) uncertainty estimation, and 3) explainability. The DR grading performance was quantitatively evaluated on several DR-labeled eye fundus image datasets, and class-wise performance was analysed. The relation between the uncertainty and the algorithm's performance was assessed. Further, the uncertainty estimation on good and bad quality images was compared, using image quality-annotated datasets, in order to assess whether bad quality images would be associated with higher uncertainties. Finally, uncertainty estimation on unfamiliar data types was also analysed, aiming at inspecting whether the estimated uncertainties could allow to detect unknown image types (*i.e.*, outliers). Different medical images were considered: colon (colonoscopy), skin, cataracts surgery and breast biopsy microscopy images. For the explainability, the attention maps produced by DR|GRADUATE were qualitatively evaluated through visual inspection, as well as quantitatively, by comparison with lesion annotations performed by an ophthalmologist.

4.3.1 Datasets

4.3.1.1 DR grading datasets

The Kaggle DR detection training dataset was used for training DR|GRADUATE, and several public and private datasets have been used for evaluation. The Kaggle DR detection, Messidor-2, IDRID, DMR and SCREEN-DR datasets, used for testing, are all described in Chapter 2, section 2.4.1. Fig. 4.5 shows the class distribution of these datasets.

As referred in that Section, the DMR dataset is labeled accordingly to the Davis scale in 4 grades (Table 2.4). The international scale grades (Table 4.1) were converted to the modified Davis scale using the following assumption: NDR = R0, SDR = R1 ∪ R2, PPDR = R3 and PDR = R4.

In the private SCREEN-DR dataset, besides the image-wise annotations, available for 966 images, 348 of these images also contain pixel-wise annotations of the following lesions: MAs, HEMs, CWSs, IRMAs, EXs, NVs, PHEMs and PFIB. MAs appear as reddish small and circular dots (Zhang *et al.*, 2010), and may occur alone or in groups (Mahendran and Dhanasekaran, 2015). HEMs are bright red spots and, contrarily to MAs, their shape and appearance has substantial variability (Kar and Maity, 2018). Exudates present a large variety of shapes, sizes and contrast (Harangi and Hajdu, 2013). EXs have a more yellowish and bright appearance and sharp

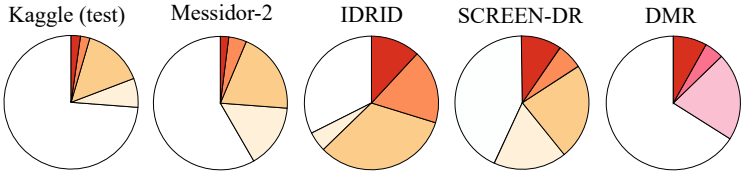


Figure 4.5: Class distribution of the DR grading datasets used for evaluation. ☐ R0 or no DR, ■ R1, ■ R2, ■ R3, ■ R4/PDR, ■ SDR, ■ PPDR.

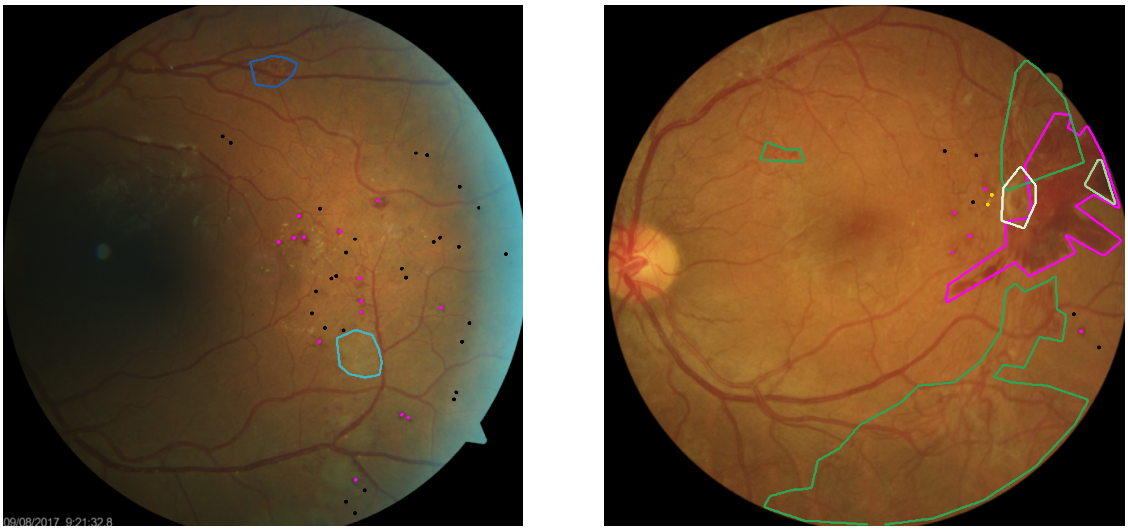


Figure 4.6: Examples of ophthalmologist's annotations on the SCREEN-DR dataset. ■ MA, ■ HEM, ■ EX, ■ CWS, ■ IrMA, ■ NV, ■ PHEM, ■ PFIB.

edges, whereas CWSs are more pale yellow/white without well defined borders (Prentasic and Loncaric, 2016). NVs are thin and weak, exhibiting irregular appearance such as tortuous patterns and loops (Gupta *et al.*, 2017). Different tools were available for annotating the images, such as dots, circles, polygons and free-hand tool, as shown in Fig. 4.6. The annotated lesions were grouped in the different grades to allow a direct correspondence with the predicted attention map for each grade. The R1 ground truth map includes only the annotated MAs; R2 contains only HEMs; R3 contains HEMs, CWSs and IRMAs; R4 includes NVs, PHEMs and PFIB. EXs were also annotated but not included in the grade maps since they are not directly evaluated in the grading process.

4.3.1.2 Quality assessment datasets

Three different eye fundus images datasets with image quality information were used on uncertainty-related experiments.

DR1: This dataset (Pires *et al.*, 2012) is constituted by 5 776 images with resolution 640×480 pixels, captured using a Topcon TRC-50X mydriatic camera (45° FOV). From these, 1 300 are labeled as being of good quality (do not contain blur and are centered on the macula) and 1 392 of poor quality (blur).

DRIMDB: The Diabetic Retinopathy Image Database (DRIMDB) (Sevik *et al.*, 2014) is composed of 216 images divided into three classes: good (125), poor (69), and outlier (22). The images were captured by a Canon CF-60UVi fundus camera (FOV of 60°), and have a resolution of 570×760 pixels. Images were annotated by an expert, and good-quality images are suitable for use in medical diagnosis by an ophthalmologist.

HRF: The High-Resolution Fundus (HRF) database (Köhler *et al.*, 2013)[3] contains 18 pairs of

[3] https://www5.cs.fau.de/research/data/fundus-images/

images from the same eye from different subjects, captured using a Canon CR-1 fundus camera (FOV of 45°). In each pair there is one poor and one good quality image. These poor quality images have decreased sharpness, which can be local or global, due to, for instance, a defocused camera. Other quality features, such as image contrast or illumination conditions, are not contemplated.

4.3.2 Evaluation

4.3.2.1 DR grading

The DR grading performance of DR|GRADUATE is evaluated using the Cohen's quadratic weighted Kappa (κ), used for measuring inter-rating agreement between raters in ordinal multi-class problems. This metric penalizes discrepancies between ratings, in a manner that depends quadratically on the distance between the prediction and the ground truth, as follows:

$$\kappa = 1 - \frac{\sum_i^C \sum_j^C w_{i,j} O_{i,j}}{\sum_i^C \sum_j^C w_{i,j} E_{i,j}} \tag{4.5}$$

where C is the number of classes, w is the weight matrix, O is the observed matrix and E the expected matrix. $w_{i,j}$ is the weight penalization for i, j element, which is given by: $(i-j)^2/(C-1)^2$. $O_{i,j}$ is the number of samples from the j^{th} class (ground truth) which are classified by the model as being from the i^{th} class. $E_{i,j}$ is the outer product between the prediction and ground truth classification histogram vectors, normalized such that $\sum_{i,j} E_{i,j} = \sum_{i,j} O_{i,j}$. κ values range from -1 (complete disagreement) to 1 (perfect agreement).

However, κ alone does not completely reflect the performance of a DR grading algorithm. When dealing with a very imbalanced problem, as DR grading (Fig. 4.5), κ is dominated by the most represented classes. Using confusion matrices allows to perform grade-wise evaluation and thus understand better the algorithm's behaviour. We computed normalized versions of the confusion matrices by dividing each value of the original matrix by the sum of the values of each row.

MIL assumption validation Preliminary experiments were performed to test whether the MIL assumption hampered the DR grading performance of the network. This was done by training a network in which the MP layer of DR|GRADUATE (Fig. 4.2) is replaced by a fully connected (FC) layer, keeping the remaining network as it was. No significant performance difference was found between the non-MIL and MIL models, with the first tending to present better R0 detection but a higher confusion among the disease cases. Because of this, no further experiments for this model were performed.

4.3.2.2 Uncertainty estimation

Experiments were performed in order to assess how the prediction performance relates with the uncertainty estimation of those predictions. In order to test whether low uncertainty images are

indeed associated with a higher performance, we defined a set of thresholds on the uncertainty values and, for each threshold, select the subset of images with estimated uncertainty below that threshold. We then compute the κ of our model predictions on these subsets. Since very low thresholds may lead to the inclusion of a number of images which do not truly represent the dataset, and the κ computation for these few samples would not be reliable, the uncertainty thresholds start at 0.15 to ensure an acceptable image subset size. This threshold was determined considering the dataset with the smaller number of images (SCREEN-DR), for which an uncertainty threshold of 0.15 lead to the inclusion of 5% of the dataset, corresponding to 48 images. Furthermore, cumulative histograms of the class-wise number of images as function of the uncertainty were produced. This allows to exclude the possibility of variations in κ being influenced by the different proportion of images per grade instead of being related with the uncertainty.

Finally, matrices of the estimated uncertainty were computed by averaging the DR|GRADUATE's produced uncertainties for all images from each position of the confusion matrix (predicted grade *vs* ground truth grade).

Sensitivity analysis A sensitivity analysis was performed to assess the image quality's influence in the uncertainty estimation. Namely, the influence of image blur was studied by smoothing the input image with Gaussian filters of increasing standard deviation values and inspecting the effect on DR|GRADUATE's uncertainty estimation. Likewise, contrast was studied since lower contrasts are also characteristic of bad quality images. For that, the image's contrast was decreased via linear histogram stretching, by setting the image's maximum intensity to a progressively lower value, and evaluating the influence on the prediction's uncertainty.

4.3.2.3 Explainability

With the purpose of evaluating the explanation map, a predicted object e_c corresponds to a connected component from the grade-wise explanation $\mathscr{E}_c^c$ after applying a fixed probability threshold. The explanation was quantitatively evaluated as follows: 1) percentage of all e_c that overlap with the ground truth map of the corresponding grade (O_{obj_g}); 2) percentage of all e_c that overlap with the ground truth map of the corresponding grade or lower (O_{obj}); 3) percentage of correctly predicted images for which the maximum activation object overlaps with the ground truth map of the corresponding grade (O_{max}); 4) percentage of correctly predicted images for which at least one e_c overlaps with the ground truth map of the corresponding grade (O_{class}); 5) percentage of ground truth objects that overlap with the predicted map of the corresponding grade or higher (O_{gt}); 6) percentage of all e_c that overlap with the ground truth map of any grade. (O_{any}). Note that the evaluations 2) and 5) are justified by the max-pooling assumption made in our pipeline: it is acceptable to assume that if a given region is detected as being from grade 2, for instance, it may also contain signs from a lower grade which are not highlighted since only the higher grade prevails after the max-pooling.

4.3.2.4 Statistical tests

Statistical tests were performed to assess significant differences between the uncertainties inferred for different datasets. Since these data samples did not follow a normal distribution, which is an assumption on most parametric statistical hypothesis tests, the non-parametric Kruskal-Wallis H Test was used (Kruskal *et al.*, 2012). Further analysing the *p-value* may not be enough when one is dealing with sufficiently large samples, since a statistical test will indicate a significant difference in most of the cases (Sullivan and Feinn, 2012), *i.e.* very large sample sizes may lead to the detection of differences that are quite small and possibly non relevant. It is thus important to report the *p-value* results but also the effect size, which measures the magnitude of the difference between variables in a statistical population. The effect size is defined as $ES = \mu_1 - \mu_2$, where μ_1 and μ_2 are the averages of the two groups. The Cohen's d was used as an effect size measure, and can be defined as follows:

$$d = ES/S \tag{4.6}$$

where S, the pooled standard deviation, is given by:

$$S = \sqrt{\frac{(n_1-1)\times s_1^2 + (n_2-1)\times s_2^2}{n_1+n_2-2}} \tag{4.7}$$

where n_1 and n_2 are the sample sizes for the two groups and s_1 and s_2 are the standard deviations for the two groups.

The interpretation of d, considering the thresholds proposed by Cohen and revised by Sawilowsky (2009), can be established as follows: 1) $d = 0.1$: very small; 2) $d = 0.2$: small; 3) $d = 0.5$: medium; 4) $d = 0.8$: large; 5) $d = 1.3$: very large; 6) $d = 2$: huge.

Having into account that our datasets are commonly on the order of more than 1 000 samples, the effect size is reported together with the significance level. Since a significant difference will most likely be achieved in these datasets, d provides a measure of the magnitude of these differences.

4.3.3 Parameter setting

The input size to the network (Fig. 4.2) is $640\times640\times3$, and the side of the output feature map, N, is 20. The receptive field, RF of the network is of 156, and the stride, s, is of 32. Regarding the attention maps generation, the σ' of $N_{2,\sigma'}$ from Eq. 4.3 was set to $RF/6$, since it allows to fit the majority of the Gaussian distribution in the kernel window. The weight α in the loss (Eq.4.2) was set to 0.7, in order to preserve the greater importance of the log-loss part but still impose some level of regularization. Regarding the class balancing, we chose r = 0.99, $w_f = (0.5, 2.0, 2.0, 3.0, 3.0)$ ($f = 300$) in Eq. 4.4, aiming at not overfitting the original dataset distribution and to increase the generalization capability of the network.

Table 4.2: Quadratic weighted kappa for DR|GRADUATE's DR grading in four different datasets. ◁ 7 025 images from the Kaggle DR detection training dataset, * 7 000 images from the Kaggle DR detection training dataset, ⋆ 4 classes, ◇ 496 out of the 9 443 images.

	Kaggle test	Kaggle	Messidor-2	IDRID	SCREEN-DR	DMR⋆
DR|GRADUATE (proposed)	0.74	0.75◁	0.71	0.84	0.74	0.78
de la Torre *et al.* (2018)	0.74	-	-	-	-	-
González-Gonzalo *et al.* (2018)	-	0.72*	-	-	-	-
Takahashi *et al.* (2017)	-	-	-	-	-	0.79◇

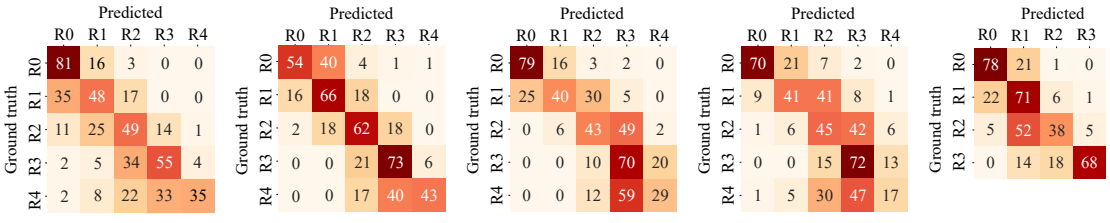
4.3.4 Training details

Images were cropped around the FOV and resized to the input size of the network, 640×640 pixels. The model was trained for 240 epochs with the Kaggle DR detection training set, using a batch size of 30, the Adam optimizer, and a learning rate of 3×10^{-4}. Experiments were performed on an Intel Core i7-5960X, 32Gb RAM, 2×GTX1080 desktop with Python 3.5, Keras 2.2 and TensorFlow 1.8.

4.4 Results

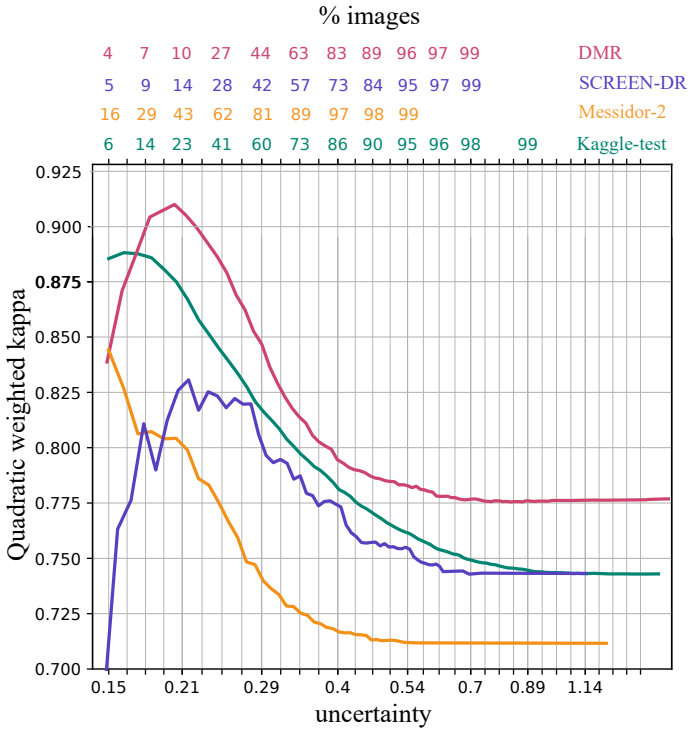
4.4.1 DR grading

DR|GRADUATE achieves similar κ to other state-of-the-art methods, as depicted in Table 4.2. Note that the majority of the state-of-the-art methods did not test on different datasets, and thus an assessment of their generalization capabilities is not possible. DR|GRADUATE was evaluated in different test datasets and the consistency of the results suggests that the system has properly learned discriminant features for DR grading. The grade-wise normalized confusion matrices (in percentage) are shown in Fig. 4.7. In general, all the matrices have a diagonal tendency, *i.e.*, the predictions are close to the ground truth. R4 is the most misclassified grade. For instance, for both the IDRID and SCREEN-DR datasets (Fig. 4.7c and 4.7d), the majority of the R4 images is confounded with R3. From the misclassifications reported in Fig. 4.7, the percentage of over-diagnosis is between 24% (Kaggle DR detection test set) and 55% (SCREEN-DR) for all datasets, which is higher than the 21% expected for human observers (Krause *et al.*, 2018). Consequently,

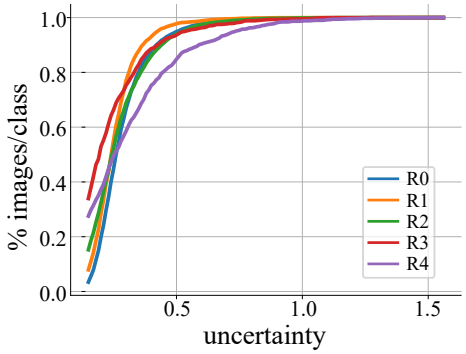


(a) Kaggle DR test **(b)** Messidor-2 dataset **(c)** IDRID dataset **(d)** SCREEN-DR **(e)** DMR dataset

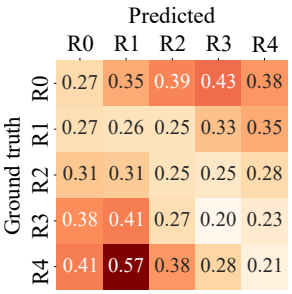
Figure 4.7: Confusion matrices of DR|GRADUATE's predictions *vs* the ground truth DR grades for different test datasets. Matrixes were normalized by dividing each value of the original confusion matrix by the sum of the values from the corresponding row (total number of images from that grade in the dataset). The shown values are in percentage.



(a) Quadratic weighted Kappa *vs* uncertainty (*xx*-axis in logarithmic scale) for four different datasets (DMR: 4 classes).



(b) Cumulative percentage of images from each grade included per uncertainty threshold, for the Kaggle DR detection test set.



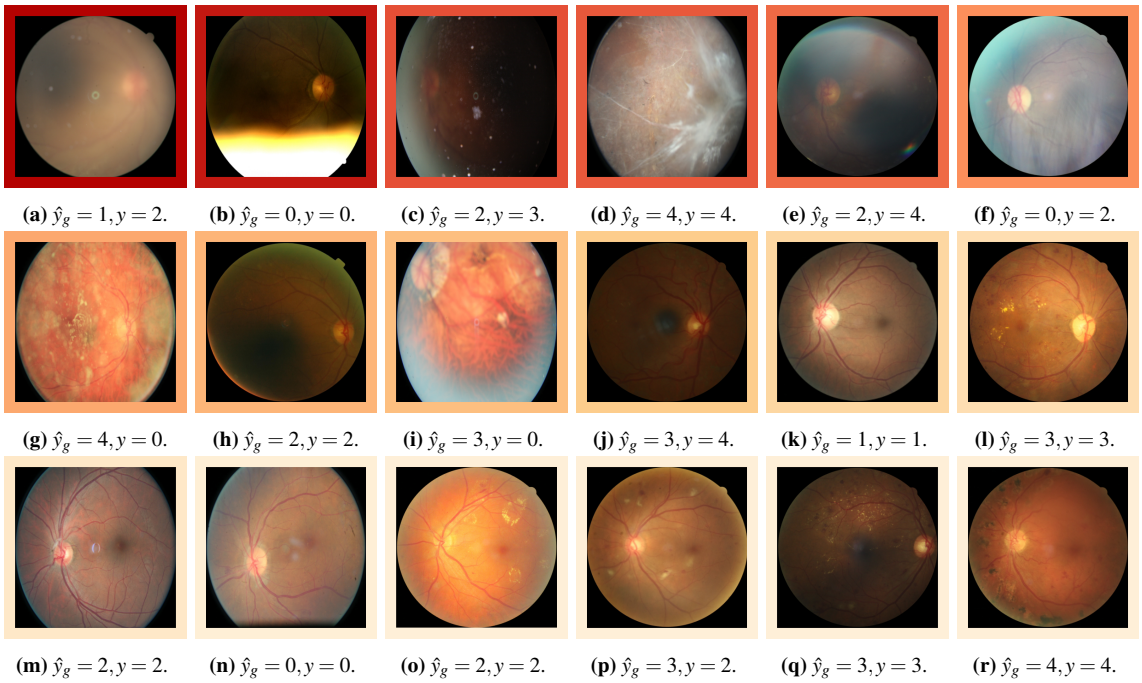
(c) Average predicted uncertainty matrix for the Kaggle DR detection test set.

Figure 4.8: Uncertainty of the predictions in relation with the DR|GRADUATE's performance.

the percentage of under-diagnosis is between 45% (SCREEN-DR) and 76% (Kaggle DR detection test set), being always lower than the 79% reported for human observers.

4.4.2 Uncertainty estimation

The evolution of κ with the uncertainty level threshold for four different datasets is shown in Fig. 4.8a. For the Kaggle DR detection test set, a κ of approximately 0.82 is achieved for the 50% lowest uncertainty images. Likewise, considering the 75% lowest uncertainty predictions, κ is of approximately 0.8. The average uncertainty matrix is shown in Fig. 4.8c, and Fig. 4.8b depicts



(a) $\hat{y}_g = 1, y = 2.$ **(b)** $\hat{y}_g = 0, y = 0.$ **(c)** $\hat{y}_g = 2, y = 3.$ **(d)** $\hat{y}_g = 4, y = 4.$ **(e)** $\hat{y}_g = 2, y = 4.$ **(f)** $\hat{y}_g = 0, y = 2.$

(g) $\hat{y}_g = 4, y = 0.$ **(h)** $\hat{y}_g = 2, y = 2.$ **(i)** $\hat{y}_g = 3, y = 0.$ **(j)** $\hat{y}_g = 3, y = 4.$ **(k)** $\hat{y}_g = 1, y = 1.$ **(l)** $\hat{y}_g = 3, y = 3.$

(m) $\hat{y}_g = 2, y = 2.$ **(n)** $\hat{y}_g = 0, y = 0.$ **(o)** $\hat{y}_g = 2, y = 2.$ **(p)** $\hat{y}_g = 3, y = 2.$ **(q)** $\hat{y}_g = 3, y = 3.$ **(r)** $\hat{y}_g = 4, y = 4.$

Figure 4.9: Uncertainty-associated predictions for images from the Kaggle DR detection test set. The color of the frame represents the uncertainty, $\hat{y}_g$ is the prediction and y the ground truth grade. Images were cropped around the field-of-view and resized to a square image. Uncertainty colorbar: 0.05 1.406

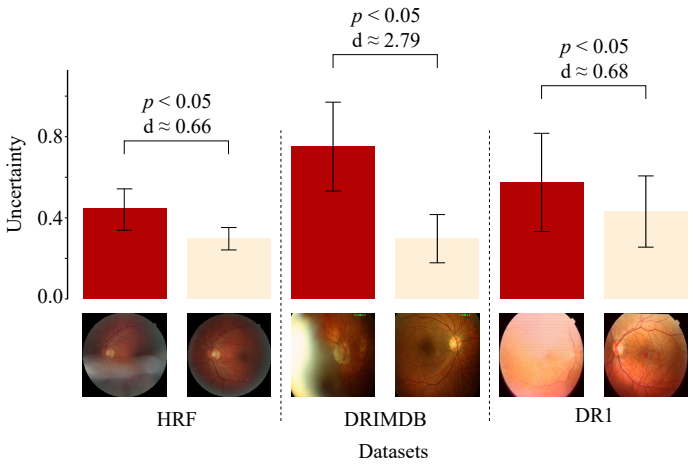
the grade-wise normalized cumulative histogram as function of the inferred uncertainty. Finally, Fig. 4.9 shows examples of Kaggle DR detection test set images along with DR|GRADUATE's uncertainties of the predictions.

The uncertainty estimation results for three quality-labeled datasets (Section 4.3.1.2) are shown in Fig. 4.10a. For all datasets, the inferred uncertainty is statistically higher for the bad quality subsets ($p\text{-}value < 0.05$), and the respective Cohen's d is huge for the DRIMDB and medium/large for the other datasets.

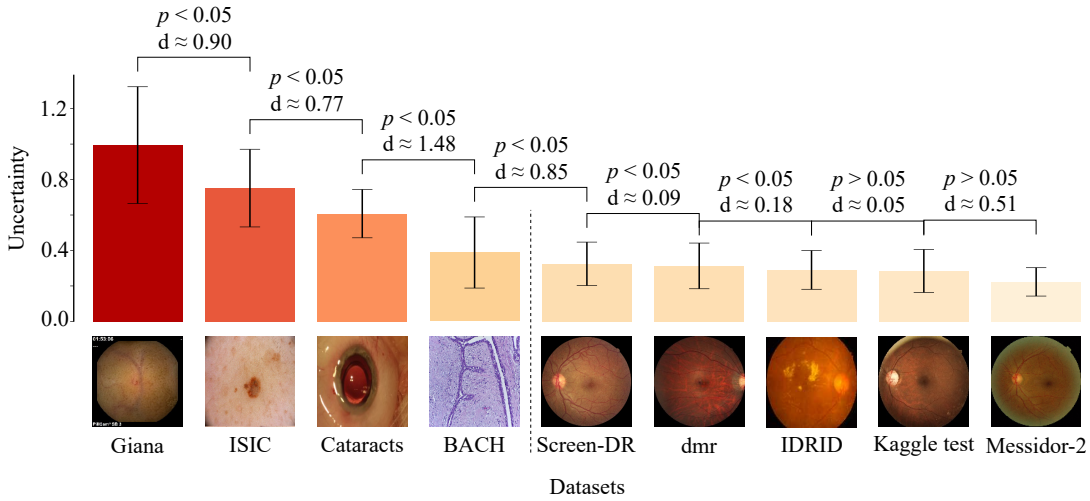
Fig. 4.10b shows the uncertainty behavior of DR|GRADUATE for datasets of unfamiliar data types along with the eye fundus image datasets used for evaluation. Between the unfamiliar data types the statistical hypothesis tests indicate that, for each dataset, the results are statistically different ($p\text{-}value < 0.05$) from the neighbouring dataset, and the Cohen's d values suggest a large to very large magnitude of the differences. From the BACH to the SCREEN-DR dataset a statistical difference is found, with Cohen's d indicating a medium magnitude of the differences. Regarding the eye fundus images, some datasets show a statistically different average uncertainty. However, in these cases, Cohen's d values indicate a small to very small magnitude of the differences.

4.4.2.1 Sensitivity analysis

Fig. 4.11 shows examples of images blurred with the Gaussian filters of increasing standard deviation (std). Fig. 4.12 depicts the results obtained for this test in 5 different datasets (Kaggle DR detection test set, Messidor-2, IDRID, SCREEN-DR and DMR), where the curve represents

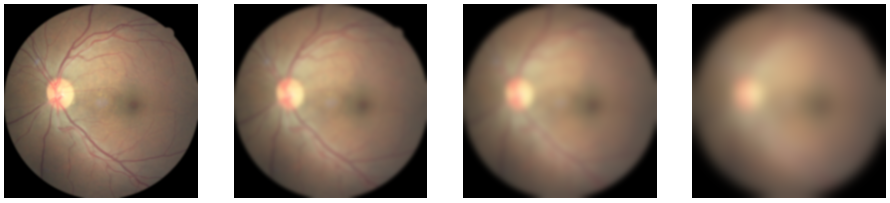


(a) Uncertainties predicted by DR|GRADUATE for good and bad quality images from public datasets. *p*-values obtained through the Kruskal-Wallis H Test and Cohen's *d* between each good and bad sets from a given dataset are presented.



(b) Uncertainties predicted by DR|GRADUATE for unfamiliar data types from different medical image datasets and for the studied eye fundus images datasets. *p*-values obtained through the Kruskal-Wallis H Test and Cohen's *d* between each dataset and its right neighbour on the plot are presented.

Figure 4.10: Uncertainty prediction for different types of tasks.



(a) $std = 3$ **(b)** $std = 6$ **(c)** $std = 10$ **(d)** $std = 20$

Figure 4.11: Examples of convolution with Gaussian filters of increasing standard deviation (*std*).

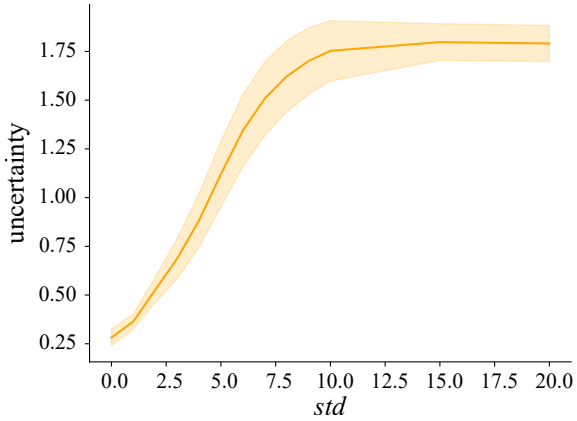


Figure 4.12: Uncertainty estimation of DR|GRADUATE for images resulting from convolution with Gaussian filters of increasing standard deviation (*std*). The uncertainty values are the average of the uncertainties for 5 datasets (Kaggle DR detection test set, Messidor-2, IDRID, SCREEN-DR and DMR) considering a 95% confidence interval.

the average of the uncertainty values for each dataset, and the shaded area corresponds to a confidence interval of 95%. The sensitivity analysis via contrast adjustment did not suggest any relation between image contrast and uncertainty estimation.

4.4.3 Explainability

Several examples of DR|GRADUATE's attention maps *vs* the specialist's annotation on SCREEN-DR dataset images are shown in Fig. 4.13. In each predicted map, the region with maximum activation is surrounded by a square. Figs. 4.13a, 4.13c, 4.13e and 4.13f illustrate examples where the maximum value of the explanation map properly explains the predicted grade. The remaining images show cases of incorrect predictions and/or incorrect explanations.

Table 4.3 shows the explanation-wise quantitative performance in terms of the number and type of highlighted lesions, O_{obj_g}, O_{obj}, O_{max}, O_{class}, O_{gt} and O_{any} (refer to Section 4.3.2.3). Fig. 4.14 shows the values of O_{max} and O_{class} obtained through the application of different thresholds to the attention maps. The higher the threshold the smaller the objects on the map, resulting in higher specificity and lower sensitivity. The achieved results suggest 0.3 to be an appropriate threshold since it allows to retain a large percentage of the lesions while producing regions not too coarse.

Table 4.3: Assessment of the produced attention maps (threshold=0.3) in comparison with specialist annotated lesions. These metrics are described in Section 4.3.2.3.

O_{obj_g}	O_{obj}	O_{max}	O_{class}	O_{gt}	O_{any}
0.506	0.677	0.712	0.784	0.526	0.729

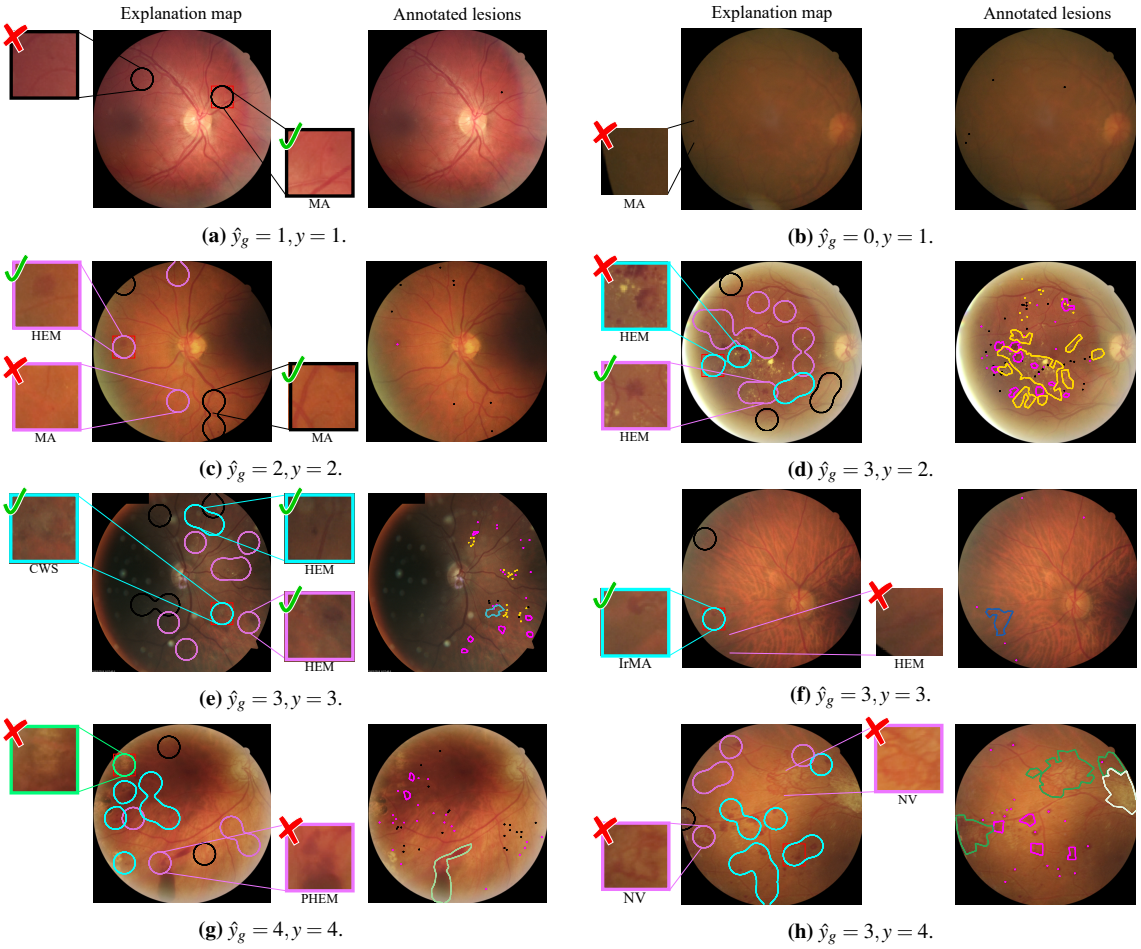


Figure 4.13: Maps predicted by our method along with the ophthalmologist's annotated lesions in the SCREEN-DR dataset. The curves are the contours of the explanation maps (threshold=0.3), and the red square indicates the region of most relevance for diagnosis (corresponding to the maximum in the network's output activation map). Close-ups of relevant regions are shown, with ✓ indicating that the region is correctly predicted, and ✗ indicating otherwise. Below each close-up the name of the corresponding ground truth lesion (if existent) is shown. Explanation map: ■ R1, ■ R2, ■ R3, ■ R4; ground truth map: ■ MA, ■ HEM, ■ EX, ■ CWS, ■ IrMA, ■ NV, ■ PHEM, ■ PFIB.

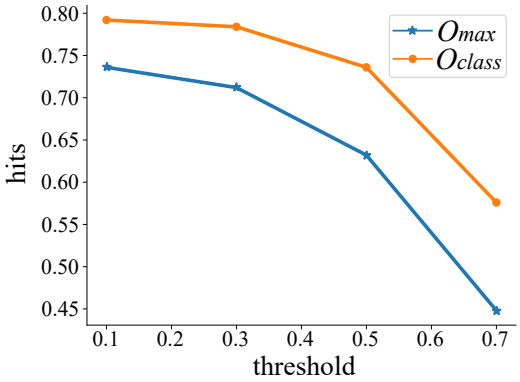
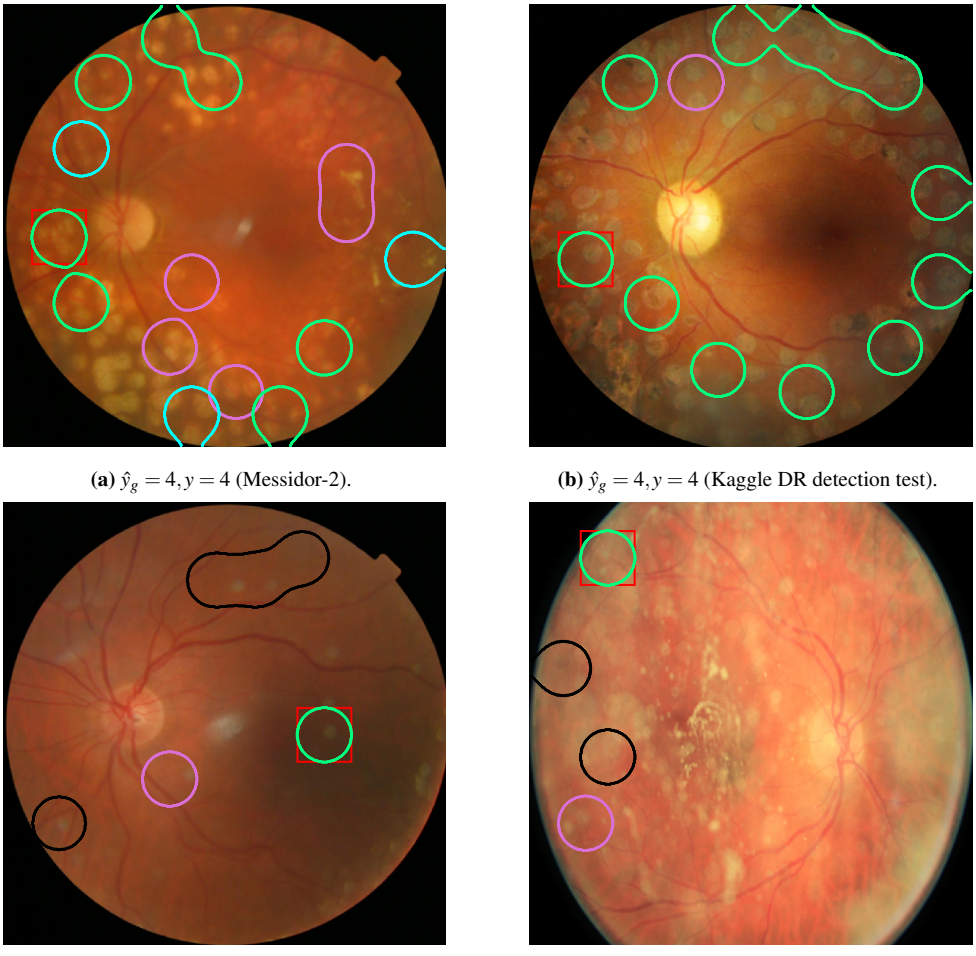


Figure 4.14: O_{max} and O_{class} for different threshold values (0.1, 0.3, 0.5 and 0.7) applied to the attention maps (Eq. 4.3).

4.5 Discussion

4.5.1 DR grading

DR|GRADUATE's performance does not degrade between the Kaggle DR detection test set, the most similar image- and grading-wise to the training data, and other datasets, showing the generalization capability of the model. Furthermore, considering the misclassified cases, the 2× higher over-diagnosis percentage of DR|GRADUATE in comparison to the expected human performance moderately promotes patient follow-up, challenging specialists to reassess their opinion. Despite the high performance achieved, DR|GRADUATE tends to fail to distinguish between R3 and R4 grades. As shown in Figs. 4.15a and 4.15b, DR|GRADUATE learned to associate photocoagulation treatment and laser marks to R4, following the training set labels. Indeed, these images may not have signs of R4, such as neovessels and pre-retinal hemorrhages, and thus should either be marked as photocoagulated and removed from the dataset or be included with the grade corresponding to the visible lesions. This issue is further exacerbated by the high class imbalance of the data (Fig. 4.5), since the data augmentation and class balancing scheme can contribute to the



(a) $\hat{y}_g = 4, y = 4$ (Messidor-2).

(b) $\hat{y}_g = 4, y = 4$ (Kaggle DR detection test).

(c) $\hat{y}_g = 4, y = 0$ (Messidor-2).

(d) $\hat{y}_g = 4, y = 0$ (Kaggle DR detection test).

Figure 4.15: Examples of images from Messidor-2 and Kaggle DR detection test sets that are predicted by DR|GRADUATE as R4. The curves are the contours of the attention maps (threshold=0.3). ■ R1, ■ R2, ■ R3, ■ R4.

regulation of the weights update but does not solve the problem of limited variability of the samples. Because of this, when inferring images with different annotation criteria or with the presence of artifacts, DR|GRADUATE misclassifies images as can be observed in Fig. 4.15c and 4.15d.

In other scenarios, misclassifications occur due to failed lesion detection or similarities between different lesion types. Namely, failures such as predicting a R1 image as R0 can be due to missing a single MA. Indeed, the small size and subtlety of these lesions make their detection particularly challenging (Fig. 4.16a, where no apparent lesion is seen) and easily confoundable with acquisition or anatomical noise (Fig. 4.16b). Confusions between R1 and R2 are likely related with the difficulty of distinguishing between the red lesions MAs and HEMs (Fig. 4.16c and 4.16d), a task that also challenges medical experts since they can be very similar in size and color. Furthermore, different grading standards also account for these classification errors. Namely, for the DMR dataset, the conversion of SDR = R1 $\cup$ R2 does not holds necessarily true in all cases. For instance, when an image is R3 by the international scale due to the presence of > 20 HEMs, it would fit in the SDR category of the Davis scale. However, through the conversion, this image would fall in the PPDR category, which can justify the model's confusion between R1 and R2 that is verified in Fig. 4.7e.

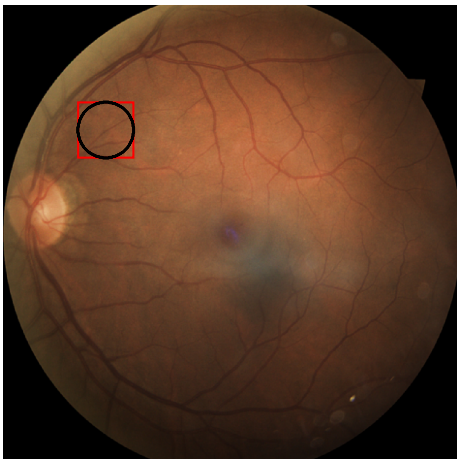
Classification errors for R2 and R3 grades are mostly due to the grading rules. Indeed, in cases where there are no other signs of R3 (Table 4.1), grading depends on HEM counting. However, this task is non-trivial for specialists since HEM may appear in clusters, hindering lesion counting and thus introducing noise on the data. Furthermore, the MIL assumption fails for these cases, as there is no novel lesion to find, leading to prediction errors (Fig. 4.17b). Instead, DR|GRADUATE may have associated lesion size to the grade. Alternatively, despite relying on a single lesion to perform grade inference, the construction of the lesion map L depends on the integration of features with different receptive fields and thus the lesion with the maximum response may indirectly encode lesion counting, serving as a proxy of the overall relevant lesion density (Fig. 4.17a).

4.5.1.1 Comparison with the state-of-the-art

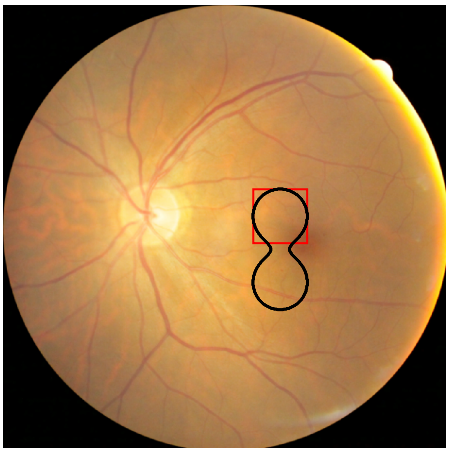
Overall, DR|GRADUATE achieves similar performance to other state-of-the-art methods and shows high generalization capability. Indeed, our model has a κ similar to Takahashi *et al.* (2017)'s method on the DMR dataset (value computed by data provided by the authors) while being trained on a different dataset and considering a different grading scale (international instead of Davis). Note that class-wise performance comparison is limited because confusion matrices are not commonly available in the literature. The proposed system achieves a performance similar to de la Torre *et al.* (2017), while having only 2% of the parameters and not requiring highly dedicated loss functions. Indeed, DR|GRADUATE's architecture already promotes the distance minimization between the ground truth and the prediction. This allows the use of the common categorical crossentropy as the main term of the loss function and thus avoids the need for the direct optimization of the κ, which requires higher computational power. On the other hand, DR|GRADUATE's performance is inferior to the 0.84 κ reported by Krause *et al.* (2018) on EyePACS-2 dataset (1 818 images) (data publicly unavailable at the time of writing). However, their model was trained on



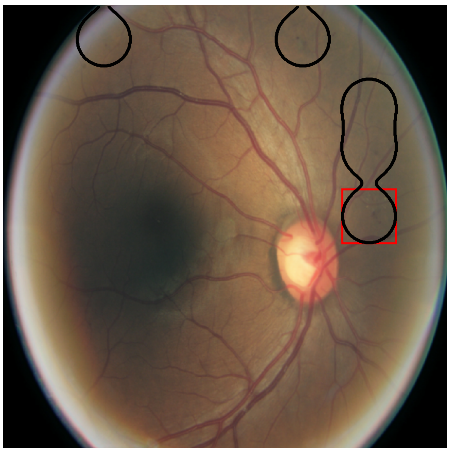
(a) $\hat{y}_g = 0, y = 1$.



(b) $\hat{y}_g = 1, y = 0$.

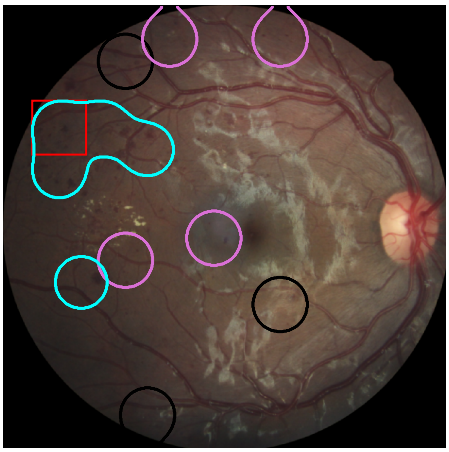


(c) $\hat{y}_g = 1, y = 2$.

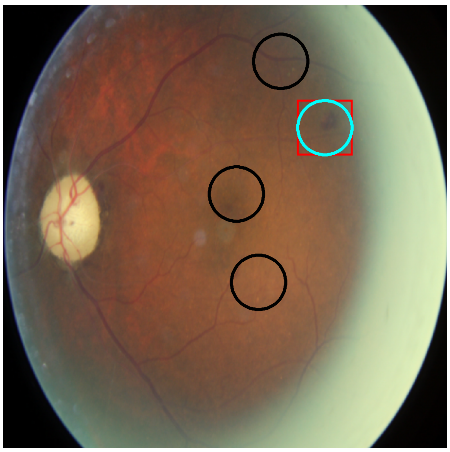


(d) $\hat{y}_g = 1, y = 2$.

Figure 4.16: Examples of R0-R2 Kaggle DR detection dataset images (not used for training) along with the DRI-GRADUATE predictions. The curves are the contours of the attention maps (threshold=0.3). ■ R1, ■ R2, ■ R3, ■ R4.

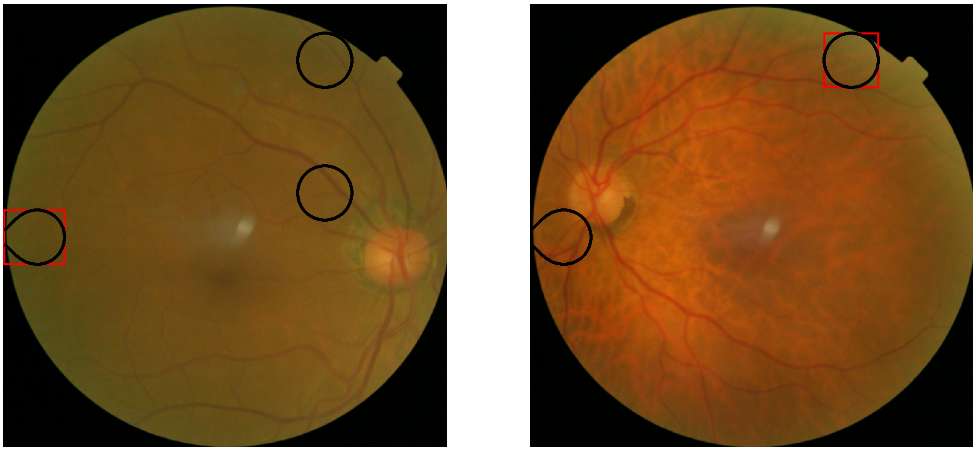


(a) $\hat{y}_g = 3, y = 3$.



(b) $\hat{y}_g = 3, y = 2$.

Figure 4.17: Examples of R2 and R3 Kaggle DR detection dataset images (not used for training) along with DRI-GRADUATE predictions. The curves are the contours of the attention maps (threshold=0.3). ■ R1, ■ R2, ■ R3, ■ R4.



(a) $\hat{y}_g = 1, y = 0$.

(b) $\hat{y}_g = 1, y = 0$.

Figure 4.18: Examples of healthy Messidor-2 images mispredicted as R1. These images show camera artifacts (present in the same location across several different images) which are detected as R1 typical lesions. The curves are the contours of the attention maps (threshold=0.3). ■ R1, ■ R2, ■ R3, ■ R4.

an expensive dataset composed of over 1.6 million images, and with ground truth labels adjudicated by multiple ophthalmologists. Further, the authors apply an empirically defined cascade of thresholds over their continuous network output in order to define the grade, ensuring that a baseline sensitivity was achieved for the higher grade classes. Consequently, the system also tends to overgrade images, with a normalized over-grading of 72%, which can lead to a high number of unnecessary follow-ups. On the other hand, DR|GRADUATE presents a lower over-estimation and thus a more moderate trade-off between over and under diagnosis of DR.

4.5.2 Uncertainty estimation

Ideally, DR|GRADUATE should infer higher uncertainties for misclassified cases. Indeed, high κ values occur for low uncertainty (Fig. 4.8a), indicating that the DR|GRADUATE's uncertainty estimation is a valid measure of the quality of the prediction. Fig. 4.8c further indicates that low uncertainty is associated with correct or close classification, since the uncertainties are considerably lower in the diagonal, and tend to increase with the class distance. Furthermore, all grades follow a similar cumulative histogram, as shown in Fig. 4.8b, indicating that the tendency illustrated in Fig. 4.8a does not result from the different proportions of each grade among the analysed image subsets. Low uncertainty of misclassifications tends to be due to the presence of lesions that greatly resemble another type, characteristic of a different grade (*e.g.*, large MAs which are similar to HEMs), or to the erroneous identification of acquisition errors as lesions (Fig. 4.18).

The lower κ for higher uncertainty cases suggests that DR|GRADUATE can be used cooperatively in the clinical practice by asking ophthalmologists to review cases that are most likely wrong. Indeed, case selection via uncertainty may significantly reduce the workload during the screening pipeline. For instance, discarding the 25% most uncertain images from Kaggle DR detection test set allows to achieve a 0.8 κ, a performance similar to human experts (Krause *et al.*,

2018), thus allowing DR|GRADUATE to be used as an independent reader capable of identifying dubious cases.

High uncertainty tends to occur in very obscured and blurred images (Fig. 4.9a, 4.9b, 4.9c, 4.9e, 4.9f). In these cases, it is not possible to make a diagnosis - at least not fully -, and thus predictions are associated with high uncertainty estimations. This will be further discussed in Section 4.5.2.1. Fig. 4.9d shows an interesting case of a R4 image correctly predicted by DR|GRADUATE but with high uncertainty (0.999). This image shows a very large PFIB (PDR sign), occupying almost half of the FOV. However, as referred, DR|GRADUATE did not correctly learn these R4 signs (even less of this size) and thus it is not expected to be certain in making a diagnosis based on such image. This large white structure may have been interpreted as partially obscuring the image and thus resulted in a high uncertainty. Further, high uncertainties can also result from the resemblance of certain structures in the images with DR signs, such as CWSs (Fig. 4.9i), and with photocoagulation marks (Fig. 4.9g). Contrastingly, DR|GRADUATE correctly predicts, with low uncertainty, the good quality and diagnosable images from Fig. 4.9k-4.9r, apart from Fig. 4.9p. We believe this case is an example of the noise present in Kaggle DR detection dataset annotations, since the image is labeled as R2, but shows several CWSs (R3 sign), leading DR|GRADUATE to predict R3 with very low uncertainty (0.05).

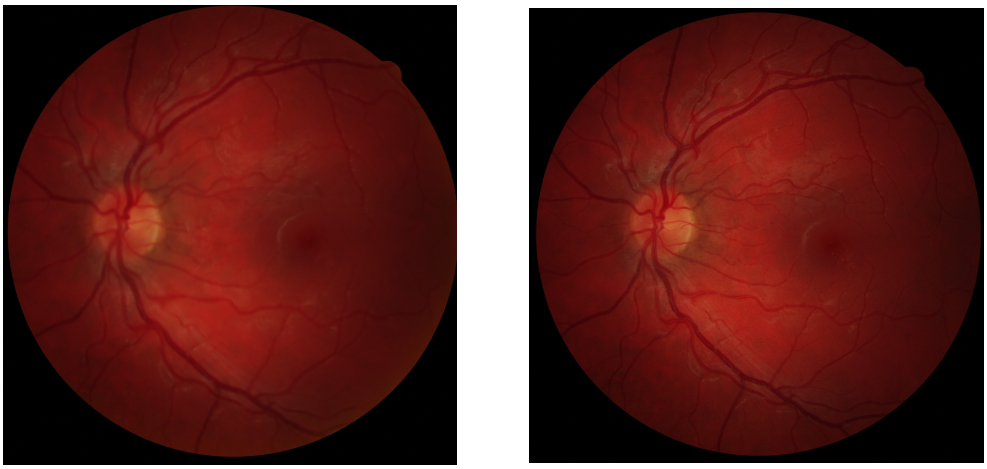
4.5.2.1 Low quality images

Poor image quality can lead to the occlusion or distortion of lesions, which hinders a meaningful and complete clinical decision (Niemeijer *et al.*, 2006). Lack of focus, sharpness or illumination often characterize bad quality images (Pires Dias *et al.*, 2014). In theory, and as suggested by Fig. 4.9, the algorithm should output higher uncertainties for images of poor quality, since in the these cases it cannot be as sure of the diagnosis as in a good quality image. Indeed, Fig. 4.9a-f mainly show examples of largely obscured images which are associated with very high uncertainty estimations. Usually, when the image is partially available for diagnosis, if DR|GRADUATE can find lesions in the viable image region it predicts a disease grade with high uncertainty (HEMs from R2, in Fig. 4.9h), whereas if it cannot find any signs it predicts it as healthy, again with high uncertainty (Fig. 4.9f).

However, since image quality is not trivial to define within the context of clinical diagnosis, image quality assessment tends to be a subjective process (Wanderley *et al.*, 2019). One can consider more than two levels of quality, such as good, partial, bad, where a partial quality image still has a region suitable for diagnosis. However, most datasets are labeled in a binary manner (good or bad quality), forcing partially obscured images to considered bad or good.

Further, as image quality is not the only factor influencing the uncertainty of the model prediction, a full correspondence between image quality and the uncertainty values is not likely. However, a tendency of bad quality images being associated, overall, with higher uncertainties is expected.

DR|GRADUATE produced significantly higher uncertainties for bad than for good quality images in each dataset (Fig. 4.10a), which suggests it can potentially be used in the DR diagnosis



(a) Bad quality image, u=0.415.

(b) Good quality image, u=0.295.

Figure 4.19: Example of a pair of good/bad quality images from the HRF dataset.

pipeline without the need for an *a priori* image quality detection step. The HRF datasets is particularly interesting since it contains pairs of good/bad quality images, but has the inconvenient of having only 18 pairs of images. Some of the HRF bad quality images differ only slightly from the good quality counterparts (Fig. 4.19), which leads to similar uncertainty estimation between the two sets. Overall, the bad quality images seem to still be suitable for diagnosis, which may justify the lower uncertainties for the bad quality subset comparing with what is verified on the other datasets. DRIMDB is the more dichotomic dataset, where images are clearly bad or good, apparently not presenting a middle term. This is reflected on the estimated uncertainties, with the bad quality set showing an average uncertainty of 0.75 ± 0.22 and the good set of 0.30 ± 0.12. Note that both HRF and DR1 contemplate only blur in the bad quality images, and thus low contrast or bad illumination are not considered.

Regarding the sensitivity analysis with the blurring operations, we can see based on Fig. 4.12 that the uncertainty shows a tendency to increase with the increase of the image blur, which decreases image quality. This corroborates our previous findings that image quality is correlated with uncertainty. When it comes to the contrast variation experiments, we believe that the uncertainty did not show any correlation with it due to data augmentation scheme performed during DR|GRADUATE's training, which contemplated brightness adjustments and contrast normalization, and most likely made the network robust to these changes. As the major factor accounting for bad quality in the tests with the bad/good quality datasets (Section 4.3.1.2 and Fig. 4.10a) was blur/decreased sharpness, not image contrast/illumination conditions, the sensitivity analysis results are according to the expectations.

4.5.2.2 Unfamiliar data types

The uncertainty estimation on the unfamiliar data types suggests that unfamiliar medical image types, *i.e.*, which are not eye fundus images, could be detected as outliers by the analysis of the

uncertainty values produced by the algorithm, which are in average higher than for the eye fundus images.

Colonoscopy images are probably the most similar to the eye fundus, showing a similar circular FOV and color (Fig. 4.10b), and sometimes even red and bright structures which may resemble retinal lesions and anatomical structures. Considering that the DR|GRADUATE was trained with eye fundus images, images that most resemble these can lead to more uncertainty since they are somehow familiar but do not present the usual and expected structures for diagnosis, similarly to what happens in bad quality eye fundus images. H&E stained microscopy images of breast tissue are the closest to the eye fundus in terms of uncertainty estimation, and are visually far from retinal images. This may be due to the presence of several small rounded structures, darker than the background (cells' nuclei) which may resemble retinal lesions. Considering that our MIL approach learns to identify the local structures responsible for a given DR grade, it makes sense that resemblance at a local level (as happens in microscopy images) leads to lower uncertainties than resemblance at the global image level (colonoscopy images).

4.5.3 Explainability

DR|GRADUATE's MIL-based architecture leads to the explanation maps to pinpoint regions relevant for diagnosis, and thus highlighting every single lesion is not expected. The detection performance is further hindered by the fact that specialists may not annotate all abnormalities (Fig. 4.13) since this can be a large time consuming task. On the other hand, DR|GRADUATE is able to correctly identify plausible regions for the reasoning behind the grade of a given image. The analysis of the produced attention maps show the ability of the system to detect lesions as small as MAs (Fig. 4.13a). As expected, when image quality is poor, image diagnosis is hindered and lesions are often missed (Fig. 4.13b). Confusion between MAs and HEMs occurs for some images (Fig. 4.13c), which is justifiable by the high resemblance of small HEMs and MAs. The detection of R3 is particularly interesting due to the variety of configurations that are associated with this grade (Table 4.1), which DR|GRADUATE seems to be able to identify: a high number of HEMs (Fig. 4.13d), some CWSs (Fig. 4.13e) or IrMAs (Fig. 4.13f). Regarding HEMs, one can see that the model detects this lesion type both as a R2 or a R3 sign (Fig. 4.13d), which is understandable since a low number of HEMs is a sign for R2 and a high number a sign of R3. Despite this image being annotated as R2, the high number of HEMs may justify the R3 prediction. However, DR|GRADUATE fails to detect NVs and other R4 lesions such as PHEMs (Fig. 4.13g) and PFIB (Fig. 4.13h). Most of the times the reason for a R4 prediction is related with the presence of photocoagulation marks (Fig. 4.13g) which corroborates the idea that these were learn by the algorithm as being R4 signs. Nevertheless, R4 lesions are sometimes highlighted in the maps as signs of disease. For instance, NVs and PHEMs are detected as R2 signs in these two images, since they slightly resemble HEMs.

Interestingly, the inherent explainability of DR|GRADUATE allows for the *a posteriori* identification of systematic errors during image acquisition. For instance, several Messidor-2 images have been wrongly classified as R1 due to the presence of dust particles in the camera lens

(Fig.4.18). However, the clear explanation provided would easily allow the clinician to correct the acquisition error and thus improve the overall quality of the screening pipeline.

4.6 Concluding remarks

We propose DR|GRADUATE, a novel deep learning-based approach for DR grading that provides an uncertainty and explanation associated with each prediction. DR|GRADUATE was trained on the Kaggle DR detection training set, and achieved state-of-the-art performance in several DR datasets. The classification of the R4 grade was the least accomplished task, suggesting the training did not correctly capture the PDR signs. This may result from the type of annotation in the Kaggle DR detection dataset, with images that underwent photocoagulation treatment and show laser marks appear to be labeled as PDR even when not presenting the typical signs of this grade.

The analysis of the predictions' uncertainties showed the tendency of lower κ for higher uncertainty cases, thus suggesting that the estimated uncertainty is a viable measure of the quality of the prediction. This means that at test time, *i.e.* where the performance of the system is not known, the uncertainty can be a measure of how much a specialist can trust the algorithm's decision, allowing to rank the predictions and select which images would benefit most from a second evaluation. Further, the explanation map produced by our algorithm should be easily interpretable by the retinal specialists, and thus constitutes a valuable tool for mitigating the black-box behaviour associated with deep NNs.

Comparing with the state-of-the-art, DR|GRADUATE has the advantage of producing both an uncertainty estimation associated with the prediction and an explanation map highlighting the most relevant regions for the classification, without compromising the DR grading performance. We believe that coupling an uncertainty estimation and explanation with the DR grade will ease the application of DR|GRADUATE in a clinical setting.

As referred, the proposed model tends to fail to detect R4 graded images. The next chapter proposes a method for mitigating the effect of the lack of amount and quality of training PDR data, via a data augmentation scheme focusing on artificial neovessel insertion on retinal images.

Chapter 5

Data augmentation for improving proliferative diabetic retinopathy detection in eye fundus images

This chapter in based on the following publication:

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., *et al.* Data Augmentation for Improving Proliferative Diabetic Retinopathy Detection in Eye Fundus Images. *IEEE Access*, 8: 182462–182474, 2020b. doi: 10.1109/ACCESS.2020.3028960

Proliferative diabetic retinopathy (PDR) is an advanced diabetic retinopathy stage, characterized by neovascularization, which leads to ocular complications and severe vision loss. However, the available DR-labeled retinal image datasets have a small representation of images of the severest DR grades, and thus there is a lack of PDR cases for training DR grading models. Additionally, the criteria for labelling these images in the publicly available datasets is not always clear, with some images which do not show typical PDR lesions being labeled as PDR due to the presence of photo-coagulation treatment and laser marks. This problem, together with the datasets' high class imbalance, leads to a limited variability of the samples, which the typical data augmentation and class balancing cannot fully mitigate. We propose a heuristic-based data augmentation scheme based on the synthesis of neovessel (NV)-like structures that compensates for the lack of PDR cases in DR-labeled datasets. The proposed neovessel generation algorithm relies on the general knowledge of common location and shape of these structures. NVs are generated and introduced in pre-existent retinal images which can then be used for enlarging deep neural networks' training sets. The data augmentation scheme was tested on multiple datasets, and allows to improve the model's capacity to detect NVs.

5.1 Introduction

The earliest diabetes-related changes in the retina are often imperceptible and have minimum impact in the vision (Acharya *et al.*, 2012), leading to the progression of the untreated disease. Contrastingly, proliferative diabetic retinopathy (PDR) is an advanced stage of DR, leading to ocular complications and severe vision loss in many developed countries (Sun *et al.*, 2018). PDR is characterized by the appearance and growth of newly formed vessels, neovessels (NV), on the retina or optic disc (OD) that extend along the retinal surface or into the vitreous cavity, increasing significantly the risk of vision loss (Sun *et al.*, 2018; The Diabetic Retinopathy Study Research Group, 1979). There is also a strong correlation between PDR and uncontrolled systemic disease (Sun *et al.*, 2018; Aiello *et al.*, 2001).

DR screening programs are thus of high importance to allow early detection and adequate treatment. Computer-aided diagnosis (CAD) systems can improve the DR screening pipeline both by reducing the burden of specialists (Scotland *et al.*, 2007) and providing a second opinion, reducing diagnosis' subjectivity (Abràmoff *et al.*, 2016; Quellec *et al.*, 2017). Deep learning has recently allowed for CAD systems to achieve near-human performance in DR detection. DR grading, *i.e.* staging of the pathology according to its severity, is a more complex task since it requires the identification and integration of different lesions. Generally, DR is classified accordingly to the international DR scale (Table 4.1, Chapter 4) as mild non-proliferative DR (NPDR) (R1), moderate NPDR (R2), severe NPDR (R3) or PDR (R4) (The Royal College of Ophthalmologists, 2012), ordered according to their risk of progression (Arévalo *et al.*, 2013). Despite the task's complexity, recent research has shown promising performance (Krause *et al.*, 2018; de la Torre *et al.*, 2017; Araújo *et al.*, 2020c). Promising results were also obtained by DR|GRADUATE, described in the previous chapter, which not only produces a grade but also an uncertainty associated with the prediction and an attention map that explains the decision. However, the proposed model tends to fail to detect R4 graded images. One of the problems with the DR|GRADUATE learning phase is the association of photo-coagulation treatment and laser marks to R4. This follows the training set labels, since training images which do not show R4 signs, such as NVs, pre-retinal hemorrhages and pre-retinal fibrosis, are still labeled as R4 due to the presence of photo-coagulation marks. The R4 detection problem is further exacerbated by the high class imbalance in the datasets, since there is a limited variability of the samples, which the data augmentation and class balancing scheme were not able to mitigate.

A possible solution to attenuate this R4-labeling and the class balancing problems is to augment the training data by generating new synthetic images from a given grade, *i.e.*, with specific lesions. Several studies have focused on the generation of artificial eye fundus images (Menti *et al.*, 2016; Costa *et al.*, 2018; Zhao *et al.*, 2018), commonly taking advantage of generative adversarial networks (GANs) to generate either an image or a portion of an image containing the structures of interest. The majority of the studies use segmented anatomical structures, namely the vasculature, to guide the generation. For instance, Costa *et al.* (2018) has proposed an end-to-end adversarial retinal image synthesis that allows to generate artificial retinal images that are anatomically con-

sistent, by training on pairs of retinal images and vessel tree segmentations. However, the method only generates healthy images, of size 256×256 pixels, which is low given the common resolution of acquired images. Although the generated vessels have an overall consistent appearance, they often present interruptions and width inconsistencies along the vessels. Indeed, the quality of the generated images may not be enough when considering the need to distinguish small structures such as DR lesions.

Other researchers have addressed the non-healthy retinal image generation. For instance, Pujitha and Sivaswamy (2018) proposed a GAN-based method to generate images with hemorrhages. The method requires vessel tree and lesion binary annotations for training the model. A new retinal image can be synthesized by providing the model the vessel tree and lesion (hemorrhage) masks. In the work of Zhao *et al.* (2018) the authors incorporated style transfer into the framework. They examined pathological cases, using as style reference images with DR and other pathologies. However, the results were not considered acceptable, with retinal images not being properly replicated.

Despite producing overall plausible retinal images, these systems present several drawbacks. For instance, they require training data with pixel-wise annotations, which are costly to obtain. Also, the training set variability determines the generation capability of the model (Costa *et al.*, 2018). Further, the high computational requirements often hinder the quality of the generated images, particularly providing low resolution images, not comparable with the higher resolutions in the retinographs. Additionally, as referred, the above developments often rely on the retinal vasculature for image generation and for evaluating its plausibility. However, the synthesized images do not present clinically accepted vessel networks (Bellemo *et al.*, 2019). Also, to our knowledge the generation of pathological images from a given DR grade, with a given lesion type, is not yet achieved (Bellemo *et al.*, 2019). The work of Pujitha and Sivaswamy (2018) suggested that it was possible to generate images with different number of hemorrhages, but it did not explore further its capability of generating images with other lesions types.

In this chapter, we propose a heuristic-based data augmentation scheme based on NV-like structures generation that compensate for the lack of PDR cases in DR-labeled datasets. The proposed neovessel synthesis algorithm relies on the general medical knowledge of common location and shape of these structures. NVs are generated and introduced in pre-existent retinal images, which can then be used for enlarging deep neural networks' training sets. Training with this type of data augmentation allows to increase detection of real NVs in independent test sets.

5.2 Methods

The NV-based data augmentation starts with the generation of prototype NVs using a semi-random graph-based approach. The resulting graphs are processed having into account the type of vessel and the location insertion, as summarized in Fig. 5.1. The processing includes moving the graph to the pixel-space, adjusting vessel width according to the branching of the structure, rotation and color matching to nearby vessels. Each step of the algorithm is detailed in this section.

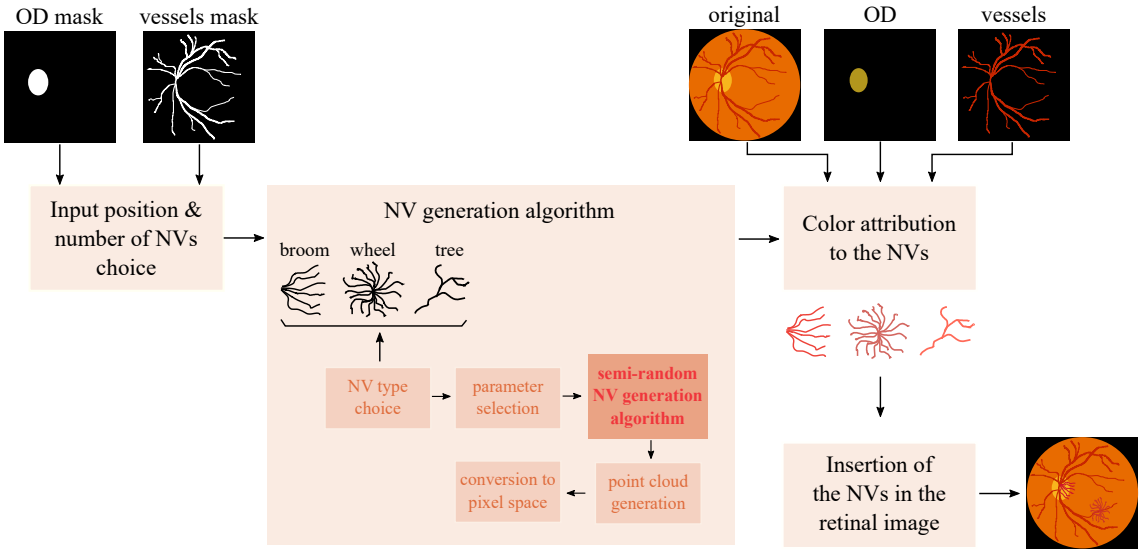


Figure 5.1: Pipeline of the proposed neovessel generation algorithm and insertion in the retinal image.

5.2.1 Insertion location

NVs occur anywhere in the retina, but they are most common posteriorly, within 45 degrees of the OD (Sun *et al.*, 2018). They are particularly common on the OD itself. In the Diabetic Retinopathy study (DRS), out of 1377 PDR eyes 15% had NV only on the OD or less than 1 OD diameter from the OD (NV on disc), 40% had NV only outside this zone (NV elsewhere), 45% had NV in both zones. NVs often lie over veins and appear to drain into them (Sun *et al.*, 2018). Sometimes NVs grow for several OD diameters across the retina without forming proeminent networks, commonly arising from the OD. In these cases, they are only distinguishable from normal vessels due to their capability of crossing both arterioles and veins (Sun *et al.*, 2018).

The region of interest (ROI) and location of where the generated NV is inserted is based on the statistical behavior of the pathology. Following the DRS statistics, the probability of inserting generated NVs only in the OD or surrounding region was set to 0.15, only elsewhere 0.40, and on both sites 0.45. The area of the ROI was set to 25% of the image size to properly capture the characteristics of the local vasculature.

5.2.1.1 Vessel segmentation

In order to ensure a more natural introduction, NVs were only inserted on pixels located on vessels. This goes in line with the fact that NV often lie over retinal veins. In order to determine which image pixels belong to vessels, a rough vessel segmentation was performed based on Otsu thresholding on the image's green channel, preceded by the application of a bottom hat operator. We call this the vessel segmented image M_{vessels}. To perform a more advanced artery/vein segmentation would be a much more complex task, which is out the scope of this study. A more sophisticated deep learning vessel segmentation method based on a U-Net was also explored in

initial tests, but the improvement in the vessel segmentation did not lead to an improvement in the performance of the DR grading algorithm, and thus the simpler approach was chosen.

5.2.1.2 Optic disc segmentation

For segmenting the OD the UOLO framework was used (Araújo *et al.*, 2018a) for the simultaneous detection and segmentation of structures of interest in medical images. This method is explained in detail in Chapter 3 of this document. It consists of an object segmentation module which intermediate abstract representations are processed and used as input for object detection. The binary image with the segmented OD is herein referred to as M_{OD}.

5.2.2 Quantity

The number of introduced NVs is determined by randomly selecting a probability threshold value. Initially the probability of inserting a NV is set to 1, and is updated after each insertion, being decreased by a factor of $\times 0.65$. The generation-insertion process continues until this probability reaches the pre-determined probability threshold and no further NVs are introduced. This update rule was selected since it allows a non deterministic choice of the number of neovessels to be inserted. The probability decay parameter value was chosen experimentally in order to allow the insertion of a plausible number of NVs, accordingly to the usual appearance of true PDR images with NVs. The most common number of inserted neovessels per image is 3 (average of 3 ± 2 NVs/image). This value seems adequate considering that 1) there is some redundancy, and having images with more than one NV may be better for training the network since it promotes activations on different regions and vessel types on the same image, which we believe that leads the network to learn more generic representations of the NVs, and 2) having a very high number of NVs would make the image non realistic.

5.2.3 Semi-random neovessel generation

The NV generation algorithm assumes that NVs are composed of branches, similarly to trees. On a tree, there is a main branch (trunk) from which several other branches can spurt. The amount, location, orientation and growth of child branches is dependent on the type of tree (or NV). By their turn, in a recursive process, the child branches can generate offspring. Generically, the algorithm operates by iteratively generating points on the Cartesian space on a parent branch. At each iteration, there is the possibility that the point belongs to a new child branch, affecting the position and distance of all sibling points thereafter. Once a growth limit has been achieved, the child branch ends and the growth of the parent branch continues.

Formally, let $b \leq b_{\max}$ define the current branch offspring level, composed by an iteratively growing set of points $P_b = \{p_0, \ldots, p_i, \ldots, p_c\}$ in the Cartesian (x,y) space. Each offspring level $b(c, L_s, \alpha, p_\alpha, D)$ has a set of properties that affect the behaviour of the branch: 1) c, with $C_b = \{c_{\min}, \ldots, c_{\max}\} \in \mathbb{N} \overset{R}{\leftarrow} c$, is the number of child branches that will offspring from the current branch of level b, randomly selected from a set of integers C_b between $c_{\min}$ and $c_{\max}$; 2) L_s, which is

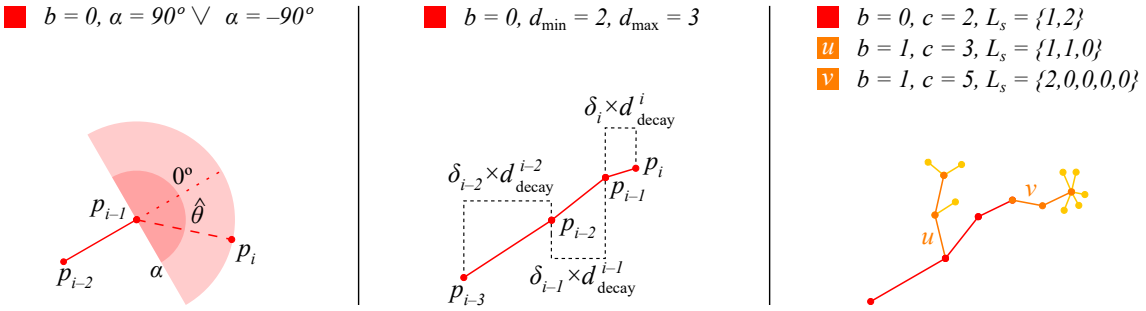


Figure 5.2: Semi-random neovessel generation algorithm's parameters example. b is the current branch offspring level, p are the points composing the branch, α is the angular range limit to where the current branch can grow, $d_{\min}$ and $d_{\max}$ are the minimum and maximum distance between two sequential points, δ is the randomly selected distance, c is the number of offspring branches and L_s are the number of points between branching offspring.

the set of number of points added between two offspring branches, with $|L_s| = c$; 3) α, the angular range limit where p_i can occur in relation to p_{i-1}; 4) p_α, which controls the probability of changing the sign of α; 5) $D = \{d_{\min}, d_{\max}, d_{\text{decay}}\}$ is a set of distance related parameters, where $d_{\min}$ and $d_{\max}$ are the minimum and maximum distances between two sequential points, respectively, and d_{decay} is a iteration-wise distance growth factor. Each point $p(\theta, \delta, s)_{b,i}$ has a behaviour inherited by the branch level b and the previous point $p_{b,i}$: 1) $\theta_{b,i} = \theta_{b,i-1} + \hat{\theta}$, where $\hat{\theta} \in [-\alpha_b, 0] \vee \hat{\theta} \in [0, \alpha_b]$, corresponding to the absolute angle in relation to the origin of the referential where p_i will be located; 2) $[d_{\min}, d_{\max}] \xleftarrow{R} \delta \times d_{\text{decay}}^{i-1}$ corresponding to the distance between $p_{b,i}$ and $p_{b,i-1}$ and; 3) s, true when $i \in L_s$, is a boolean indicating if the point will originate a new child branch. Branch growth can be done using different update rules. In our approach, points p can result from a linear coordinate update ($x_i = x_{i-1} + cos(\theta) \times \delta$, $y_i = y_{i-1} + sin(\theta) \times \delta$) or from a logarithm spiral-like curve update ($x = ae^{k\theta} \cos\theta \times \delta$, $y = ae^{k\theta} \sin\theta \times \delta$, where a and k are pre-defined constants). Due to the recursive nature of the method, the maximum number of branching levels b, b_{max}, has to be limited, otherwise the vessels would grow indefinitely. An exemplification of some of the main parameters is shown in Fig. 5.2. Illustrative step-wise examples of NVs generation are shown in Fig. 5.3. Changing the hyperparameters that control the range of possible orientations, distances, number of points and number of branches allows to obtain different NV types.

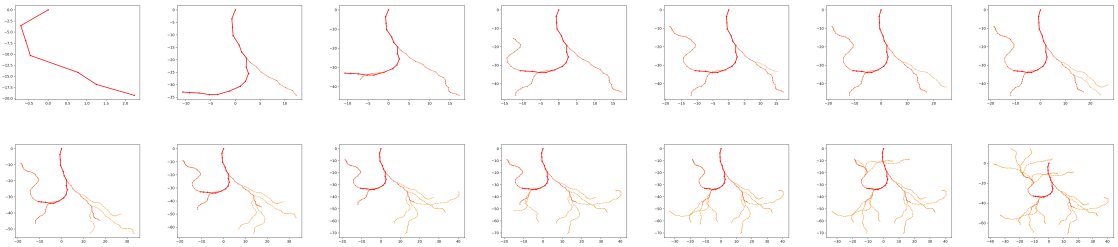
The generated NV structures are iteratively converted to the pixel space starting from the lowest branch level b. First, a randomly selected scale factor (between 2 and 6) is applied to the generated NVs in order to increase the range of possible sizes of these structures. Then, for each b, the coordinates of the points are rounded to the nearest integer. The resulting structure is dilated using as structuring element a disk of radius $\propto 1/b$ and smoothed with a Gaussian filter. The insertion of the pixel represented NV, NV_{im}, on the eye fundus image is detailed in Section 5.2.4.

5.2.3.1 Shape

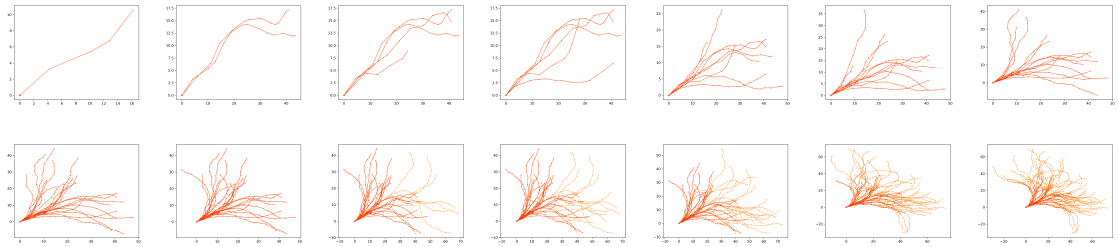
New vessels often resemble carriage wheels, radiating from the center of the structure to a circular boundary. They may also be irregularly shaped and not present this radiating pattern. Initially,

Table 5.1: Ranges and values of the semi-random generation algorithm parameters for each of the considered neovessel shapes. All shapes have three levels ($b \in \{0,1,2\}$). The coordinate update rule is always linear, except for the wheel-like shape, for which the spiral update rule is used in the last level ($b = 2$)). Also, $p_\alpha = 0.5$, except for the wheel at $b = 2$ where $p_\alpha = 0$ to promote curling. $d_{decay} = 0.95$, except for the wheel where $d_{decay} = 1$.

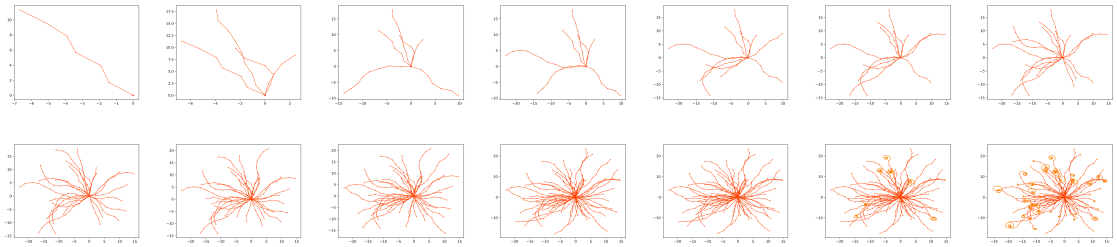
NV type	parameter - level	c *#child branches*	Ls *#points between branches*	α *angular range limit*	d *distance between points*
tree	$b=0$ $b=1$ $b=2$	$[3,4]$ $[4,5]$ 1	relatively low # branches; each level has >childs than the previous	$\in\{4,5,6\}$ $\in\{4,5,6\}$ $\in\{4,5,6\}$	$\pi/6$ $\pi/6$ $\pi/6$	grow in a more or less specific direction (not too spread)	$[3,5]$ $[1.2,3]$ $[1.2,3]$	bigger branches in the beginning than in the end (for the same *Ls*)
broom	$b=0$ $b=1$ $b=2$	$[15,25]$ $[2,3]$ 1	number of branches in the first level for high density	$\in\{0\}$ $\in\{4,5,6\}$ $\in\{4,5,6\}$	in the first level all branches come from the same point	$\pi/7$ $\pi/7$ $\pi/7$	grow in a more or less specific direction (not too spread)	- $[2,7]$ $[2,5]$	
wheel	$b=0$ $b=1$ $b=2$	$[50,60]$ $[1,2]$ 1	high number of branches in the first level for high density appearance	$\in\{0\}$ $\in\{4,5,6\}$ $\in\{75,100\}$	large # of points in the last level to allow a complete spiral	2π $\pi/6$ $\pi/5$	branches may spread in all directions in the beginning (dense)	- $[2,2.5]$ $[1.5,2]$	



(a) Tree-like shape.



(b) Broom-like shape.



(c) Wheel-like shape.

Figure 5.3: Examples of step-wise generation of artificial neovessels for the different considered shapes.

NVs may be barely visible, but later their caliber is commonly 1/8-1/4 that of a major retinal vein at the disc margin, and occasionally be as large as these veins (Sun *et al.*, 2018).

Despite NVs may present a high variety of shapes, we considered three different main types of shapes: tree-like, wheel-like and broom-like. The first shape resembles a simple tree, with relatively light branching. The broom shape has a more dense appearance and grows mainly in a given angular range. The wheel shape differs from the broom since it can grow in the full 2π *rad* angular range and its branches may end in curly structures. Different sets of parameter values were selected to associate with each distinct NV shape. These three main shapes are shown in Fig. 5.3. Table 5.1 presents a summary of the values/ranges of the most relevant parameters for each of the considered NV shapes, along with a brief explanation of some of the choices. For instance, in the case of the wheel-like shape, the parameter c of the first level ($b = 0$), which controls the number of offspring branches, must be very large in order to produce a dense structure, and thus c_{min} and c_{max} must be high. Also, the α parameter, which controls the angular range limit for branch growth, must be very large (around 2π) in order to allow the growth in all directions. Contrarily to the wheel-shape, in the tree and broom-like shapes the angular range for the first level must be reduced in order to force a more narrow angular growth, in a specific direction. The broom shape differs from the tree one mainly due to the setting of the c parameter, which must be higher for the

broom in the beginning of the growth to allow a more dense structure. Also, in the first level of the broom shape (as well as in the wheel shape) all the branches should grow from the same point, *i.e.*, L_s is has just one index, whereas in the tree they may arise from different locations.

5.2.3.2 Orientation

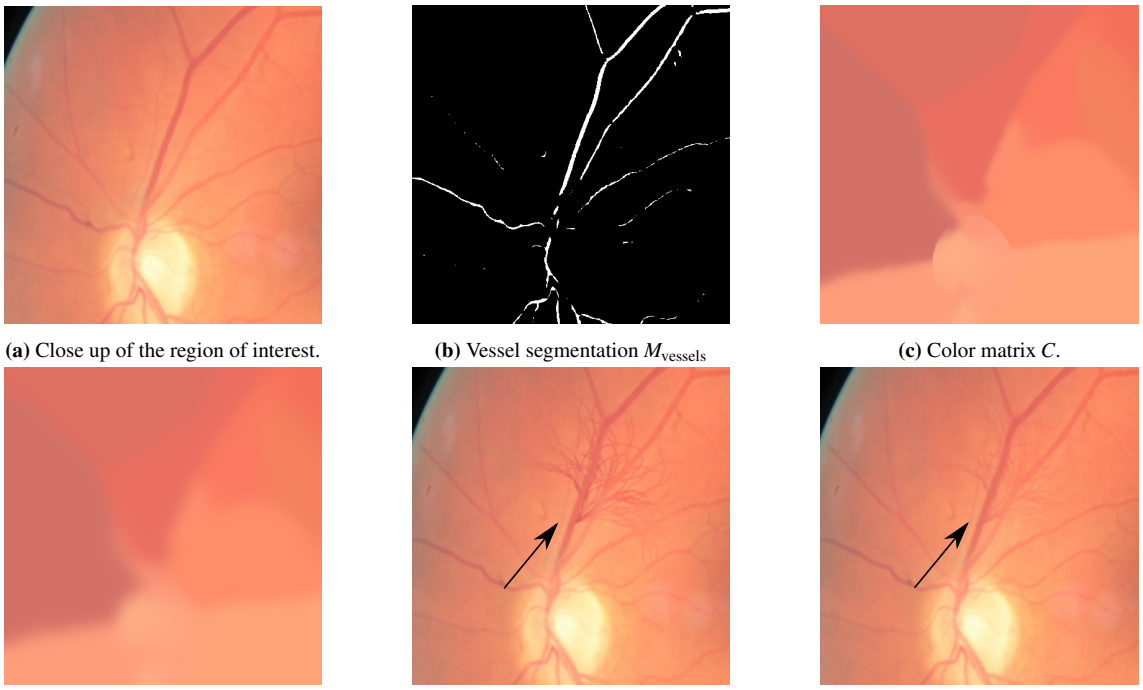
For the broom-like shape, the orientation of the NV in the image has special importance since, for instance, it may seem unnatural to have a NV orientated perpendicular to the existing vasculature. A rough indicator of the orientation of the main vessels in a given image region can be determined based on the OD position, with the definition of quadrants centered on the OD. A broom NV is thus inserted with an orientation close to that of the vessels in the insertion quadrant ($\pi/4$, $3\pi/4$, $-3\pi/4$, or $-\pi/4$), with a tolerance of 0.17 rad (randomly determined for each insertion).

5.2.4 Color assignment and neovessel insertion

The color of the generated NVs must be coherent with the ROI's vasculature to allow a realistic insertion. This color is determined based on the assumption that 1) vessels' colors are not constant and change according to multiple factors, including vessel type (artery or vein), proximity to the OD or fovea, thickness, external illumination and position on the eye; 2) the artificial NVs should have a color similar to the local vasculature, namely the nearest vessels. Because of this, the color of the NV is based on the color of the vessels from M_{vessels} near the insertion region, instead of using only the color of the vessel from which the NV will branch. Further, since the vessel segmentation is not perfect, larger vessels, expected to be more representative of the overall vasculature, are given more weight on the final NV color.

Vessel color is attributed via a color matrix C that is computed based on the colors of the local vasculature as exemplified in Fig. 5.4a. Specifically, each pixel from C, $C(i,j)$, results from the weighted average of the median RGB values of each of the objects from M_{vessels} (Fig. 5.4b). For instance, for the red channel, $C(i,j)_{\text{red}} = \sum_{v=0}^{N} r_{v,AP} \times DT_v(i,j) \times c_{v,\text{red}}$, where v is the index of each of the N vessels segmented in M_{vessels} (*i.e.*, each connected component), $c_{v,\text{red}}$ is the median red color of vessel v, $r_{v,AP}$ is the area/perimeter ratio and $DT_v(i,j) = DT'_v(i,j) / \sum_{v=0}^{N} DT'_v(i,j)$, where $DT'_v \in [0,1]$ is the complement of the distance transform of $\overline{M_{\text{vessels},v}}$. The area/perimeter ratio constitutes an approximation of vessel thickness, given that the vessel shape can be approximated by a rectangle with length c and thickness, l, with $A_v = c \times l$ and $P_v = 2(c+l)$, $r_{AP,v}$ should be higher for thicker vessels (if $c >> l \rightarrow r_{AP,v} \propto l$). Due to the high color difference that exists between the OD and non-OD regions, the vessels from each of these regions are handled separately. The color matrices are then combined (Fig. 5.4c) and C is smoothed with a Gaussian filter to reduce the influence of sharp color transitions (Fig. 5.4d).

Each pixel of NV_{im} is attributed the color of the corresponding position in the color matrix C. To ensure that NV_{im} is properly blended with the image and that branches fade with increasing distance to the insertion point and neighbor vessels, NV_{im} is weighted by the complement of the



(a) Close up of the region of interest.

(b) Vessel segmentation M_{vessels}

(c) Color matrix C.

(d) Color matrix C after Gaussian smoothing.

(e) Inserted NV without the distance transform ponderation.

(f) Inserted NV with the distance transform ponderation.

Figure 5.4: Color attribution to a generated neovessel (NV), which origin is indicated by the arrow.

normalized distance transform of $\overline{M_{\text{vessels}}}$ (see Fig. 5.4e and 5.4f). The insertion of the NV is then achieved via $\text{ROI}_{\text{with NV}} = (1 - NV_{\text{im}}) \times \text{ROI} + NV_{\text{im}} \times C$.

5.2.5 DR Grading

5.2.5.1 Method - DR|GRADUATE

DR grading was obtained using the methodology proposed in Araújo *et al.* (2020c), described in detail in the Chapter 4 of this document. DR|GRADUATE consists in a deep learning-based DR grading computer-aided diagnosis system that supports its decision by providing an interpretable explanation and an estimation of the uncertainty of that prediction, which gives the ophthalmologist a measure of how much that decision can be trusted.

The model showed a robust performance for DR grading. However, as referred in Section 5.1, it usually fails to detect R4/PDR labeled images.

5.2.5.2 Newly proposed data augmentation

DR|GRADUATE was herein trained with the original Kaggle DR detection images for grades R0-R3, and the R4 class was augmented using the images with artificial NVs. Fig. 5.5 shows examples of Kaggle images originally from grades R0-R3 with the generated NVs. Examples of newly generated images were presented to an ophthalmologist, who assessed their appearance and confirmed that the generated NVs resemble true neovascularizations. The original R4 images were

Table 5.2: Quadratic weighted kappa (κ) for the new and initial (DR|GRADUATE's) DR grading in the SCREEN-DR, Kaggle DR detection test and Messidor-2 datasets.

Dataset	SCREEN-DR	Kaggle test	Messidor-2
κ - initial	0.74	0.74	0.71
κ - new	0.78	0.74	0.70

kept, and the same number of newly generated images was added. NVs were inserted in images equitably pooled from the set of R0-R3 images. No insertion was performed on the pre-existant R4 images since this could impair the detection of other PDR signs. The reason behind this R4 training set construction was to have a significant number of images with generated NVs, but still maintain the information regarding the original data in order to not degrade the performance in the Kaggle DR detection and other datasets which may present the same labeling scheme.

The training scheme was kept the same as in the previous chapter, in order to assess only the influence of the NV augmentation in the results. Note that the original data augmentation scheme, used in Araújo *et al.* (2020c) (flips, rotations, brightness adjustments and contrast normalization), were still performed in the new training, together with the newly proposed NV augmentation.

5.3 Results

5.3.1 Datasets

The datasets herein used are the Kaggle DR detection and Messidor-2 (public) and the SCREEN-DR (private). These datasets are described in Chapter 2, Section 2.4.1. In sum, Kaggle DR detection has 35 000 in the training set, 55 000 in the test set, and the distribution of the images per grade in this dataset is {73.5, 7.0, 15.0, 2.5, 2.0}% for {R0, R1, R2, R3, R4} grades, respectively. Messidor-2 has 1 748 images, with a class distribution of {58.3, 15.5, 19.9, 4.3, 2.0}%. Finally, SCREEN-DR has 966 images, from which 348 have pixel-wise DR lesion annotations. The class distribution in this dataset is {43.0, 17.6, 23.4, 6.2, 9.8}%.

5.3.2 Results and discussion

The quadratic-weighted Cohen's kappa (κ) values of the new and of the initial DR|GRADUATE predictions for three different datasets are shown in Table 5.2. Additionally, in order to perform a grade-wise evaluation, confusion matrices of the initial and new predictions were computed, as shown in Fig. 5.6. These are normalized versions of the confusion matrices, obtained by dividing each value of the original matrix by the sum of the values of each row.

As shown, the κ values remain approximately the same in the public datasets (Kaggle and Messidor-2), and improved from 0.71 to 0.74 in the SCREEN-DR dataset. By inspecting the confusion matrices one can see that the grade-wise performance in the Kaggle dataset remained approximately the same. The fact that the performance in this dataset did not degrade suggests that the inclusion of the original R4 images together with the newly generated allowed the model to be

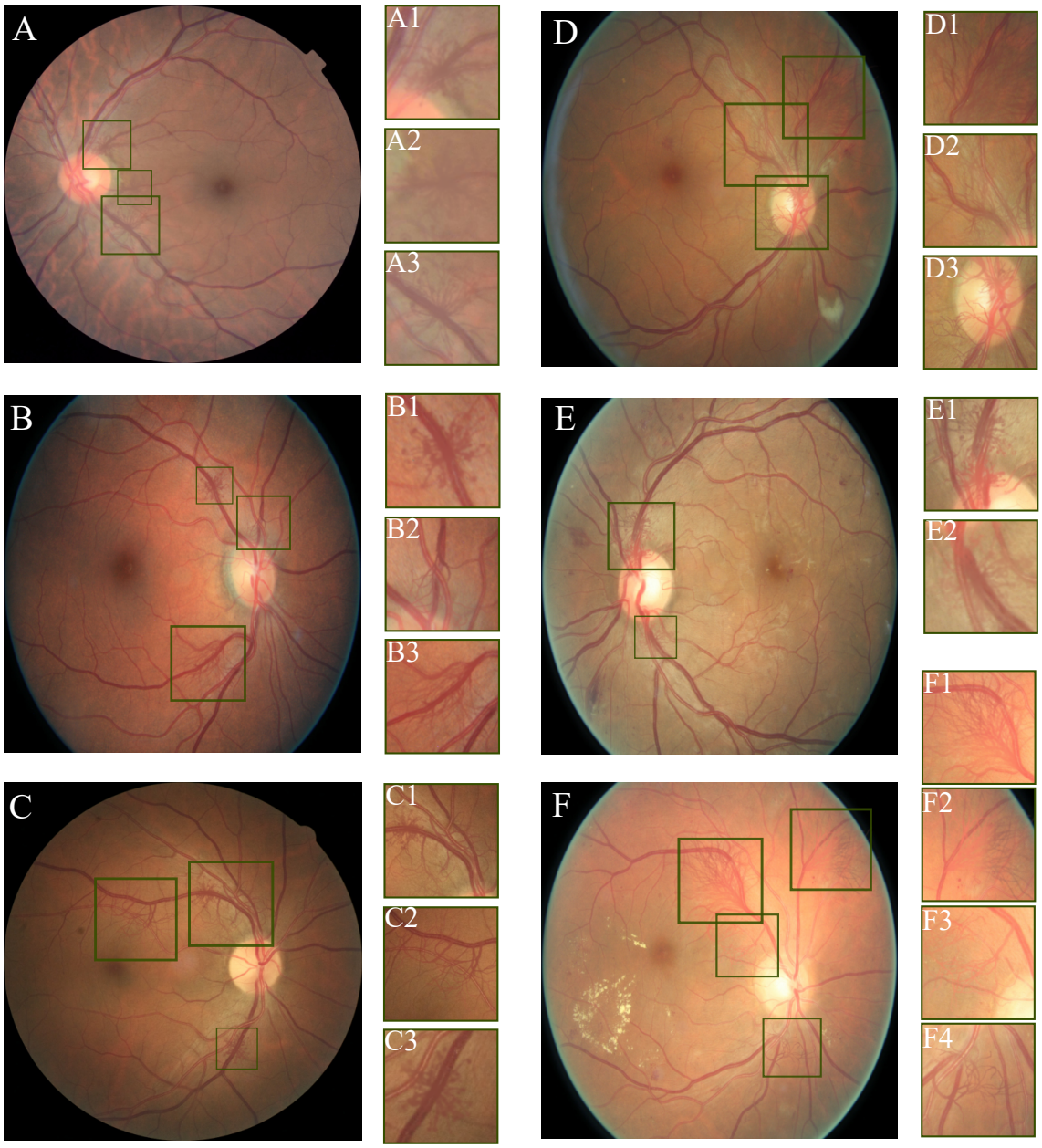
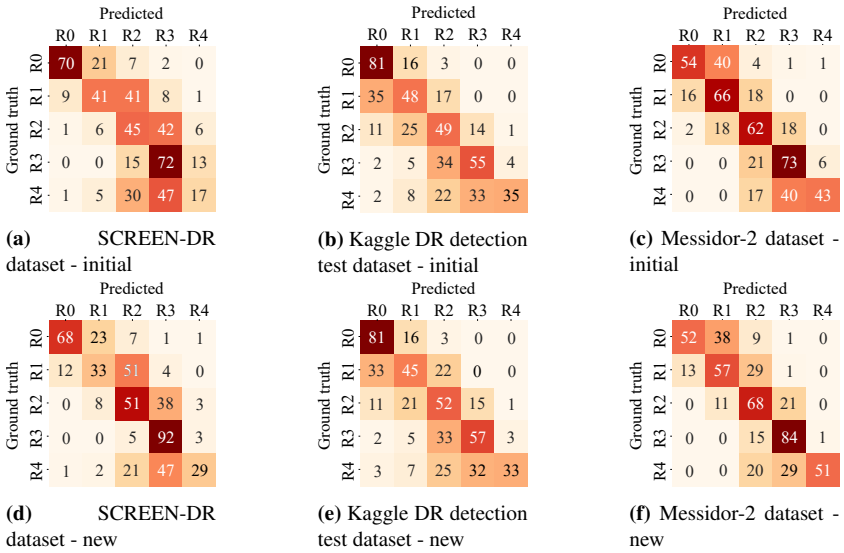


Figure 5.5: Examples of images with synthesized artificial neovessels (NVs). Close-ups of some of these NVs are shown next to the images, along with the NV type. Original DR labels (from the Kaggle DR detection dataset): A: R0; B: R0; C: R1; D: R2; E: R3; F: R3.



(a) SCREEN-DR dataset - initial

(b) Kaggle DR detection test dataset - initial

(c) Messidor-2 dataset - initial

(d) SCREEN-DR dataset - new

(e) Kaggle DR detection test dataset - new

(f) Messidor-2 dataset - new

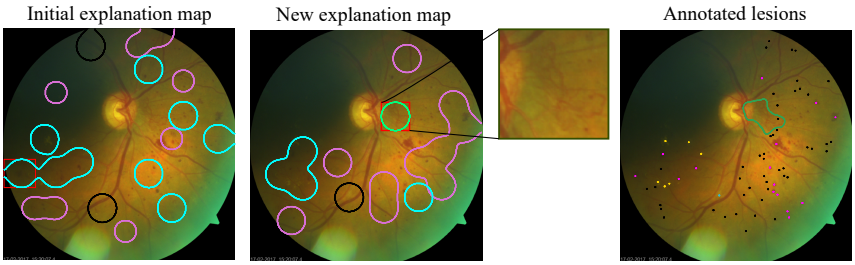
Figure 5.6: Confusion matrices of the new and initial (DR|GRADUATE's) predictions *vs* the ground truth DR grades in the SCREEN-DR, Kaggle DR detection test and Messidor-2 datasets. Matrices were normalized by dividing each original value by the sum of the values from it's row (*i.e.*, number of images from that grade in the ground truth), and the shown values are percentages.

able to classify in a similar manner the original R4-labeled images. For Messidor-2, despite the κ value remaining the same, one can see that the performance on the R4 class, both in this dataset and in the SCREEN-DR, has improved, with more R4-labeled images being correctly detected.

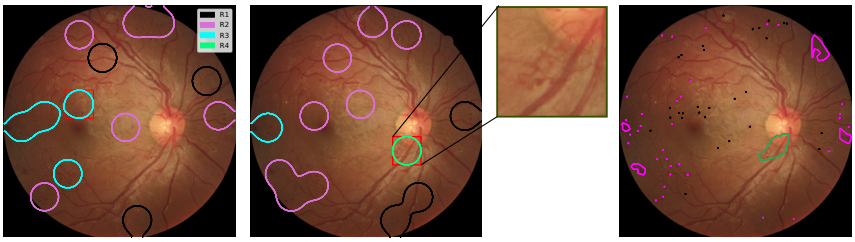
Examples of the new and of the initial DR|GRADUATE's attention maps *vs* the specialist's annotations on SCREEN-DR images are shown in Fig. 5.7. In each predicted map, the region with maximum activation is identified by a surrounding square. The model is now capable of detecting NVs in eye fundus images, as shown in Fig. 5.7a, 5.7b and 5.7c, which were initially not being detected by DR|GRADUATE (Araújo *et al.*, 2020c). However, in some cases other structures are confounded with NVs and mistakenly detected as PDR signs (Fig. 5.7c). Additionally, NVs are sometimes misdetected as other lesions, such as IRMAs (Fig. 5.7d), characteristic of R3. Note that DR|GRADUATE originally was confounding the NV in this image with an HEM rather than an IRMA. The difficulty for specialists to distinguish between neovessels and IRMAs is widely referred in the literature, being specially true for early NVs (Sun *et al.*, 2018).

One should note that PDR images, which do not present NVs but rather other PDR signs (PHEMs and PFIBs), may still be misclassified since these lesions are not contemplated in the proposed data augmentation scheme. However, one can see in Fig. 5.7e that despite not being able to identify the PHEMs in the image, the method identified one of them as being an hemorrhage, which is understandable due to the similarity HEMs and these small PHEMs.

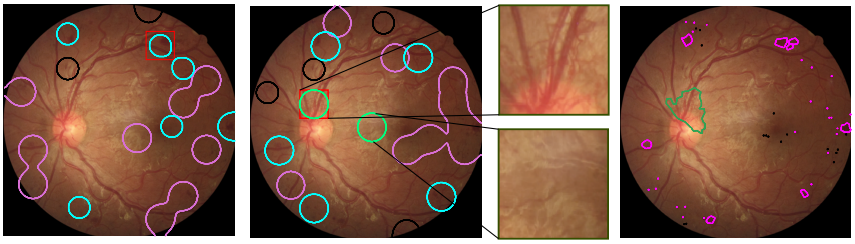
A few NVs are still missed, as the cases shown in Fig. 5.8. Some of these vessels may present an unusual shape and appearance (Fig. 5.8a and 5.8b). NVs that are too slight were missed in certain cases (Fig. 5.8c), despite being contemplated in the NV generation (Fig. 5.5)-A2). Very large NVs, such as the ones shown from Fig. 5.8d, may be too large for the network to detect given its receptive field. The NV attributed colors at times may not be very realistic, which mainly



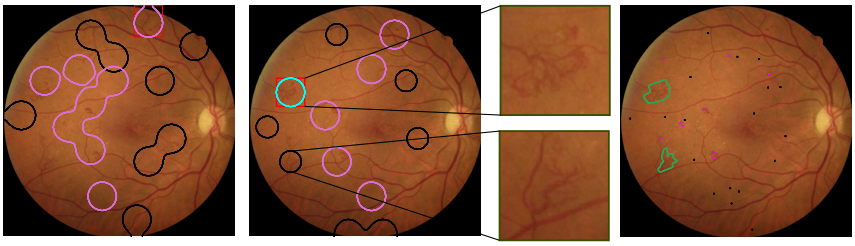
(a) Initial $\hat{y}_g = 3$, New $\hat{y}_g = 4$, $y = 4$.



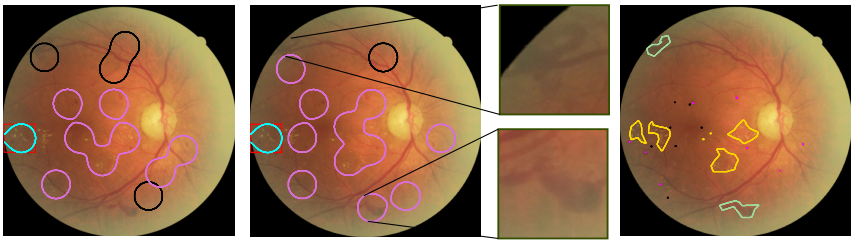
(b) Initial $\hat{y}_g = 3$, New $\hat{y}_g = 4$, $y = 4$.



(c) Initial $\hat{y}_g = 3$, New $\hat{y}_g = 4$, $y = 4$.

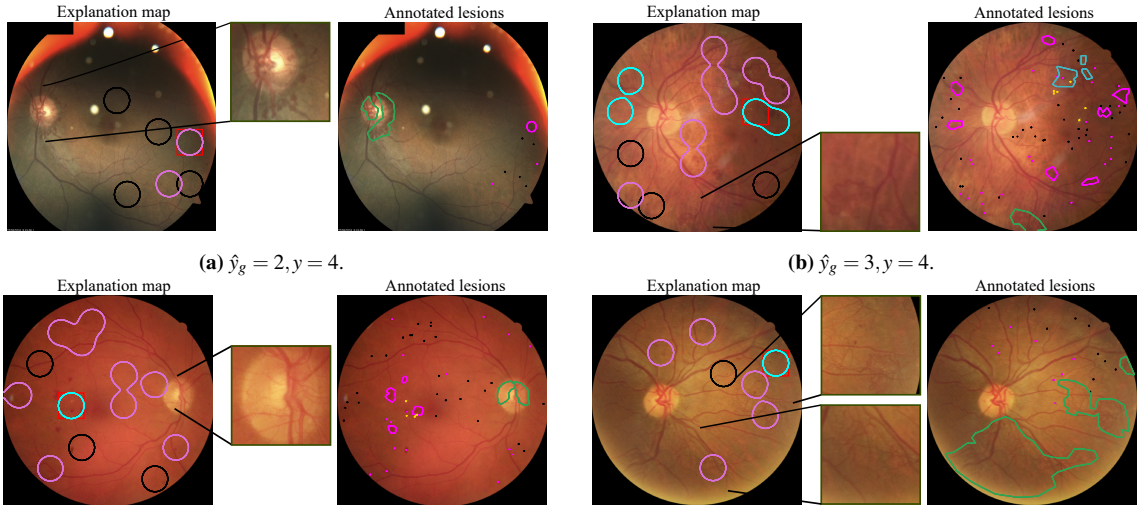


(d) Initial $\hat{y}_g = 2$, New $\hat{y}_g = 3$, $y = 4$.



(e) Initial $\hat{y}_g = 3$, New $\hat{y}_g = 3$, $y = 4$.

Figure 5.7: Maps predicted by our method and DR|GRADUATE along with the ophthalmologist's annotated lesions in the SCREEN-DR dataset. Curves correspond to the contours of the explanation maps (threshold=0.3), and the red square indicates the region of most relevance for diagnosis (corresponding to the maximum in the network's output activation map). Explanation map: ■ R1, ■ R2, ■ R3, ■ R4; ground truth map: ■ MA, ■ HEM, ■ EX, ■ CWS, ■ IrMA, ■ NV, ■ PHEM, ■ PFIB.



(a) $\hat{y}_g = 2, y = 4$. **(b)** $\hat{y}_g = 3, y = 4$.

(c) $\hat{y}_g = 3, y = 4$. **(d)** $\hat{y}_g = 3, y = 4$.

Figure 5.8: Maps predicted by our method along with the ophthalmologist's annotated lesions in the SCREEN-DR dataset in cases in which the model mispredicts R4-labeled images. Curves correspond to the contours of the explanation maps (threshold=0.3), and the red square indicates the region of most relevance for diagnosis (corresponding to the maximum in the network's output activation map). Explanation map: ■ R1, ■ R2, ■ R3, ■ R4; ground truth map: ■ MA, ■ HEM, ■ EX, ■ CWS, ■ IrMA, ■ NV, ■ PHEM, ■ PFIB.

happens due to the presence of light artifacts, given rise to color aberrations, as shown in Fig. 5.5-F2 and F3.

5.4 Concluding remarks

We proposed a data augmentation scheme to compensate for the lack of PDR cases in DR-labeled datasets. It builds upon a heuristic-based algorithm for the generation of neovessel-like structures which relies on the general knowledge of common location and shape of these structures. The synthesized NVs can be introduced in pre-existent retinal images which can then be used for enlarging the datasets for training deep neural networks. The insertion of the NVs in the retinal images has in account the color coherence with the neighbour vasculature to ensure a realistic insertion. Experiments were performed to assess the influence of this data augmentation scheme in the training of a previous proposed model for DR grading, DR|GRADUATE. Results have shown that the PDR detection performance has improved, with NVs that were being missed by the model now being highlighted in the explanation maps. However, part of the PDR images are still not detected since they do not present NVs but rather pre-retinal fibrosis or pre-retinal hemorrhages, which were poorly learned by the model. NVs with an unusual shape or that are too slight are still being missed by the model, likely due to its lack of representation in the generated dataset. This study shows the potential of introducing NVs in retinal images for improving the detection of these proliferative DR signs, thus allowing to improve the performance of computer-aided DR grading methods and easing their clinical application.

Chapter 6

Analysis of the performance of specialists and an automatic algorithm in diabetic retinopathy grading

In this chapter, we present a scheme for estimating the DR grade of an image based on an ensemble of the methodologies described in the previous chapters. Then, the final method is evaluated on different public datasets, and compared with recent state-of-the-art methods. Finally, a thorough analysis on the private SCREEN-DR dataset is performed, both of the automatic algorithm and of the specialists involved in the annotation of the dataset.

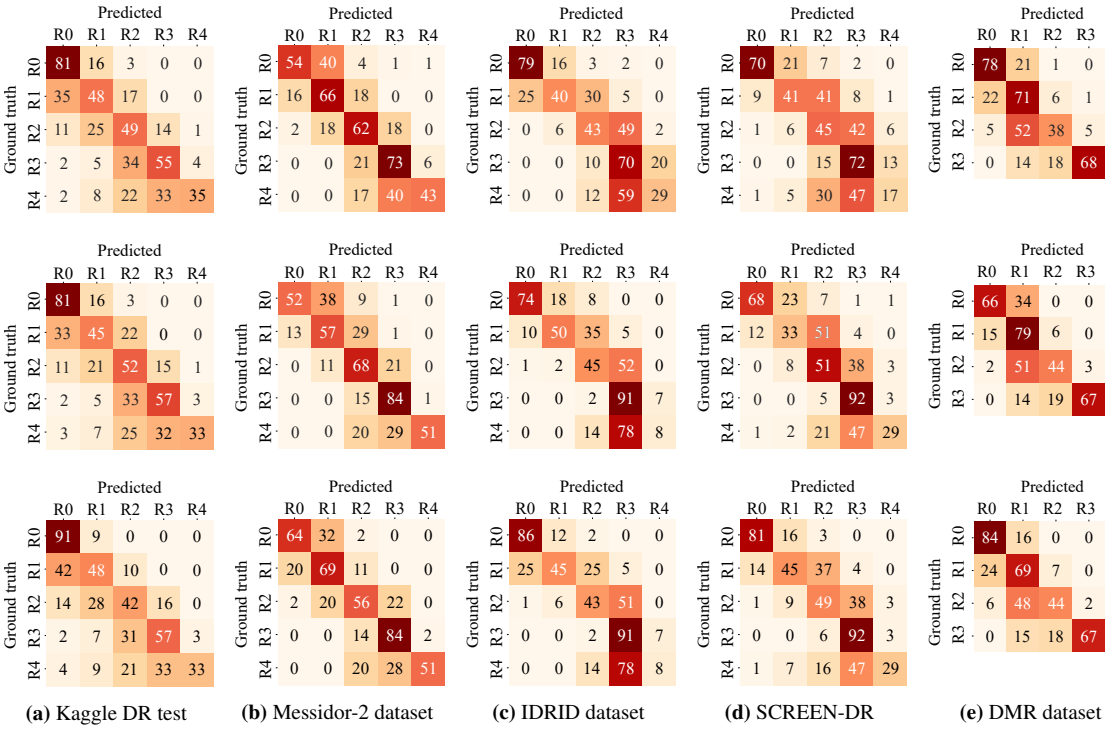
6.1 DR grading prediction

As described in Chapters 4 and 5, two models were trained for performing DR grading. The first, the original DR|GRADUATE model, $\mathscr{M}_{orig}$, was trained with the Kaggle DR detection dataset, whereas the second, $\mathscr{M}_{NV}$, was trained with an augmented version of this dataset, in which we artificially inserted structures that mimic neovessels, that are typical proliferative DR lesions. The purpose of this scheme was to mitigate the lack of representative PDR samples in the existent datasets for training, and thus improve the PDR detection performance of the network.

$\mathscr{M}_{NV}$ showed better performance in detecting higher grades than $\mathscr{M}_{orig}$, but in some cases deteriorated slightly its performance in detecting the lower grades - healthy (R0) and mild NPDR (R1), as can be seen in Chapter 5, Fig. 5.6. Based on this, an ensemble of the two trained networks, $\mathscr{M}_{orig}$ and $\mathscr{M}_{NV}$, was built in order to obtain a final DR prediction (p), using the following rules:

- p_{final} = R0 if $p_{\mathscr{M}_{orig}} =$ R0 or $p_{\mathscr{M}_{NV}} =$ R0;
- p_{final} = R1 if $p_{\mathscr{M}_{orig}} =$ R1 and $p_{\mathscr{M}_{NV}} \neq$ R0, and
- $p_{final} = p_{\mathscr{M}_{NV}}$ in the remaining cases.

In sum, the predictions of $\mathscr{M}_{NV}$ were kept, except for the cases in which $\mathscr{M}_{orig}$ predicts R0 or R1. In the latter case, the R1 prediction is kept only if $\mathscr{M}_{NV}$ did not predict R0. In the end, the



(a) Kaggle DR test **(b)** Messidor-2 dataset **(c)** IDRID dataset **(d)** SCREEN-DR **(e)** DMR dataset

Figure 6.1: Confusion matrices of DR|GRADUATE's predictions *vs* the ground truth DR grades for different test datasets. First row: original predictions; second row: trained with neovessel-augmented dataset; third row: ensemble. Matrices were normalized by dividing each value of the original confusion matrix by the sum of the values from the corresponding row (total number of images from that grade in the dataset), and the shown values are percentages.

prediction is R0 if either of the models predicts R0. Indeed, the ensemble results have shown that each of the networks detected some R0 cases not detected by the other, and thus the union of these predictions is beneficial. As can be seen in the confusion matrices from the previous chapters (herein repeated in Fig. 6.1) the networks are particularly good at detecting R0, which is by a large margin the most represented class in the training dataset, and thus it makes sense to trust these predictions.

As the uncertainties produced by the two networks are not in the same range, they were calibrated in order to allow a proper uncertainty estimation by the ensemble. The uncertainties u estimated by $\mathscr{M}_{orig}$ and $\mathscr{M}_{NV}$ on the Kaggle DR detection training set were sorted (for each of the models) and a linear regression was applied to the pairs of points $(u_{\mathscr{M}_{orig}}, u_{\mathscr{M}_{NV}})$. The resulting slope ($m = 1.5$) was used for dividing the uncertainties of $\mathscr{M}_{NV}$.

6.2 Performance evaluation

The performance of the ensemble on the different datasets is shown in Fig. 6.1, in terms of normalized confusion matrices. In general, we can see that there is an increase of the performance of the ensemble in comparison with the original models. As discussed in Chapter 5, the improvement in the detection of PDR cases from $\mathscr{M}_{orig}$ to $\mathscr{M}_{NV}$ is a consequence of the augmentation of the training dataset with the artificially generated NVs. This improvement is particularly visible in

Table 6.1: Quadratic weighted kappa for DR|GRADUATE's DR grading in four different datasets. * 7000 images from the Kaggle DR detection training dataset, ⋆ 4 classes, ⋄ 496 out of the 9443 images.

	Kaggle test	Kaggle	Messidor-2	IDRID	SCREEN-DR	DMR⋆
DR|GRADUATE (ensemble)	0.78	-	0.77	0.87	0.82	0.80
de la Torre *et al.* (2018)	0.74	-	-	-	-	-
González-Gonzalo *et al.* (2018)	-	0.72*	-	-	-	-
Takahashi *et al.* (2017)	-	-	-	-	-	0.79⋄
Gonzalez-Gonzalo *et al.* (2020)	0.77-0.80	-	-	-	-	-
Galdran *et al.* (2020b)	0.79	-	0.80	-	-	-

the SCREEN-DR private dataset, where the criteria for annotating the images is well known and where there are significant examples of PDR images with proeminent NVs. A deeper analysis on the SCREEN-DR dataset will be performed later in the following section. The exception is the IDRID dataset, for which the PDR detection performance actually decreased from $\mathscr{M}_{orig}$ to $\mathscr{M}_{NV}$, and consequently in the ensemble. The R4 annotated images from this dataset indeed show several cases with pre-retinal fibrosis and pre-retinal hermorrhages, and only a small portion have easily identifiable NVs. As discussed, these signs are not recognized by the network, and since $\mathscr{M}_{NV}$ learned the association of NVs with PDR, the network's capability for detection the R4 cases from IDRID is reduced.

In Table 6.1 the obtained κ values are compared with other relevant studies in the field, and Fig. 6.2 shows the normalized confusion matrices from the respective studies. Please refer to Chapter 2 for a description of these methodologies. Besides the κ, we computed new evaluation metrics, as in Galdran *et al.* (2020b), namely, the mean Area Under the Receiver-Operator Curve in its multi-class version, considering each possible class pair (Hand and Till, 2001). Another computed metric is the Average of Classification Accuracy (ACA), which corresponds to the mean of the diagonal values in the normalized confusion matrix (Guo *et al.*, 2019). Additionally, the Kendall τ coefficient (Kendall, 1938) was computed, which measures the ordinal association between two quantities, being a measure of the correspondence between two rankings. Values close to 1 indicate strong agreement, whilst values close to -1 indicate strong disagreement. Table 6.2 shows these results for the methods for which these metrics were available or were computable based on the information provided in the respective publications.

Table 6.1 shows that all methods present similar κ, ranging from 0.74 to 0.80 in the Kaggle DR detection test set. Note that all these methods are deep learning-based, published on 2017 or later, and most account for the ordinality of the DR grading problem in some manner. Nevertheless, only ours and that of Gonzalez-Gonzalo also propose an explanation map for the algorithm's decision. However, the approach from (Gonzalez-Gonzalo *et al.*, 2020) highlights all lesions in a single map whereas ours proposes a *per* grade explanation. Additionally, DR|GRADUATE is the only that also presents an uncertainty associated with the diagnosis. As the majority of the studies evaluate their results only on the Kaggle DR detection dataset it is not possible to infer on generalization capability of these approaches.

The confusion matrices show additional information about the different approaches for DR

Table 6.2: Performance of published methods for DR grading. mAUC: area under the curve of the multi-class version of the Receiver-Operatior Curve; ACA: Average of Classification Accuracy; *tested on a held-out portion of the Kaggle DR detection composed of 1560 images.

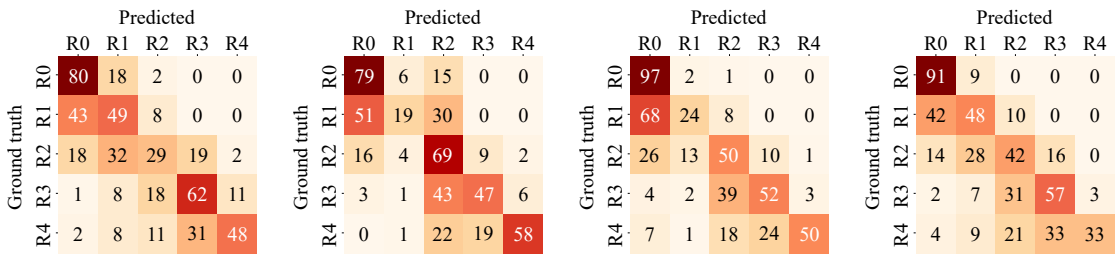
	Kaggle test	Messidor-2
	mAUC / ACA / Kendall-τ	mAUC / ACA / Kendall-τ
DR|GRADUATE (ensemble)	0.88/ 0.54 / 0.71	0.90 / 0.65 / 0.68
de la Torre *et al.* (2018)	- / 0.54 / -	-
Zhao *et al.* (2019)*	.../ 0.54 / -	-
Galdran *et al.* (2020b)	0.89 / 0.55 / 0.71	- / 0.63 / -

grading. For instance, DR|GRADUATE and Galdran *et al.* (2020b) show the most diagonal tendency, and also the best sensitivity for healthy cases (R0). Galdran *et al.* (2020b) shows, however, poorer performance in R1 detection, predicting 68% of the true R1 images as healthy. Zhao *et al.* (2019) seems to be less successful in detection the R1 grade, showing a tendency to either predict these images as healthy or R2. DR|GRADUATE is the one that performs the worse for R4. Nevertheless, and as discussed in the previous chapters, the annotation of the Kaggle's R4 images is not the most reliable, and thus the performance of R4 detection in this dataset is not a good indicative of the performance of the method in real-world practice.

6.3 SCREEN-DR analysis

In this section, we analyse the performance of the ensemble in our private SCREEN-DR dataset, as well as of the four specialists that were involved in the annotation of this dataset. The main objectives of this analysis are the following:

- establish a gold standard of inter-observer variability in DR diagnosis in order to allow a fair comparison with automatic methods;
- compare the developed method for DR grading with the annotators performance in the SCREEN-DR dataset;



(a) de la Torre *et al.* (2018) **(b)** Zhao *et al.* (2019) **(c)** Galdran *et al.* (2020b) **(d)** DR|GRADUATE.

Figure 6.2: Confusion matrices of published DR grading methods in the Kaggle test set (the method from Zhao *et al.* (2019) was validated on 1560 images from the Kaggle training set). Matrices were normalized by dividing each value of the original confusion matrix by the sum of the values from the corresponding row (total number of images from that grade in the dataset), and the shown values are percentages.

- explore possible relations between annotation data and between annotation data and the algorithms' prediction (uncertainty *vs* image quality, uncertainty *vs* annotation time, among others).

6.3.1 Available annotation data

The SCREEN-DR retinal images were collected in the context of a Portuguese DR screening program, managed by the Portuguese North Health Administration (ARSN). These were acquired using Canon CR-2 AF and Carl Zeiss Meditec AG VISUCAM PRO NM nonmydriatic retinal cameras and have different resolutions (with width and height ranging approximately from 1 500 to 2600 pixels). Images were annotated in a web platform (Pedrosa *et al.*, 2018) by independent annotators. Among other, the following information is available: 1) image quality: bad, partial or good quality; 2) DR grading (when the image is diagnosable): R0, R1, R2, R3 or R4; 3) time stamps of the button clicks; 4) annotator ID (herein named *538a*, *d56a*, *c10f* and *f28b*). Four specialists annotated the images, not all having annotated the same data. A total of 1 217 images was labeled with a DR grade, with 440 having been labeled by a single annotator, 199 images by two annotators, 122 by three, and 456 by all four annotators. Please note that up until this section, *i.e.*, all previous chapters and Section 6.2, the performance on the SCREEN-DR analysis has always been performed considering as ground truth the annotator *538a*.

6.3.2 Intra-observer analysis

Fig. 6.3 shows the annotation times of each annotator *vs* the attributed DR grade. One cannot identify a consistent pattern relating annotation time and the annotated DR grade. However, for some annotators, there is a slight tendency of spending more time annotating images that they labeled as having higher grades. This can be due to the fact that, in some cases, images of higher grades present more signs of disease, and thus have more content to be analysed by the annotator.

Fig. 6.4 plots the percentage of partial quality labeled images *vs* the annotated image grade. Note that for the images which the annotator was able to provide a DR grade there are only two possible quality labels: partial or good. There is no clear relation between both, with image quality labels being approximately evenly distributed among the DR grades.

6.3.3 Inter-observer analysis

We defined confidence of an annotation as the number of annotators that agree with the most commonly attributed grade, *i.e.*, obtained by majority voting. For a given number of annotators, N, the confidence can range from 1 to N. Fig. 6.5 shows the distribution of annotated images for each of the confidence levels, separately for the images with 2, 3 or 4 annotators. Results suggest that most of the times N or $N-1$ annotators agree. When four specialists annotated the images, there is still a significant number of images for which only two of them agreed.

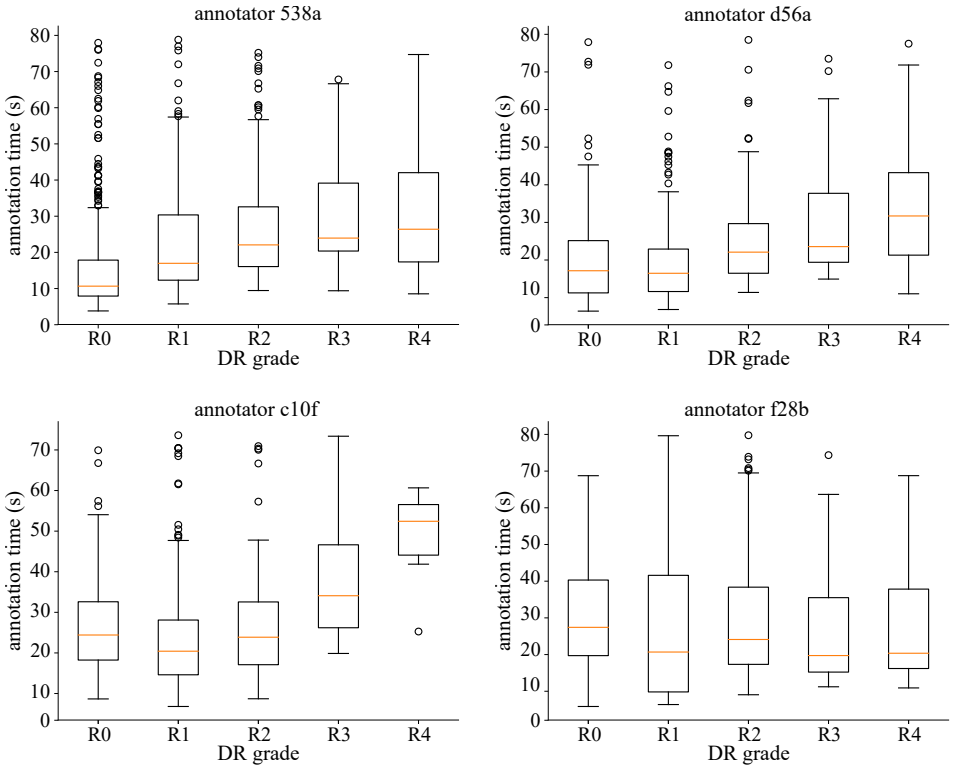


Figure 6.3: Annotation times (in seconds) vs DR grades.

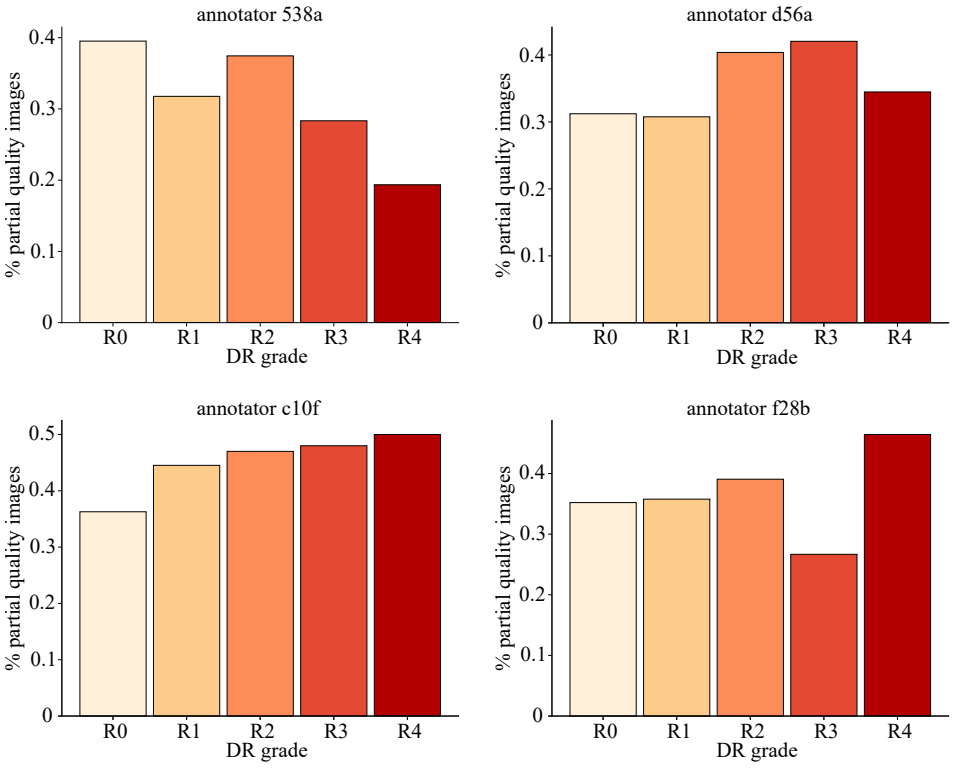


Figure 6.4: Percentage of partial quality images (out of the partial and good quality images) accordingly to the attributed DR grade.

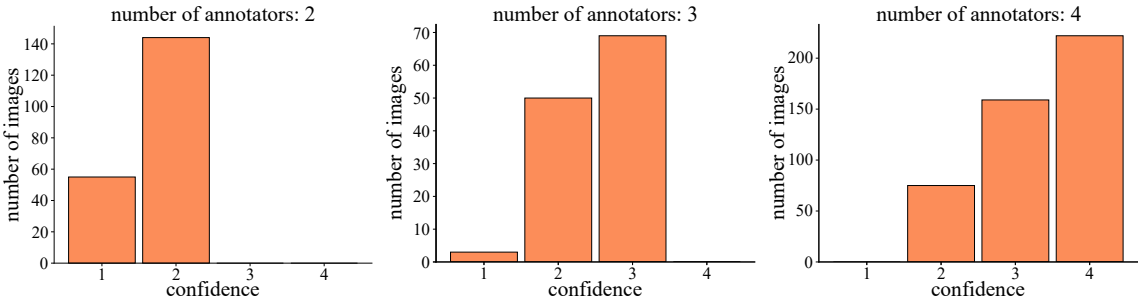
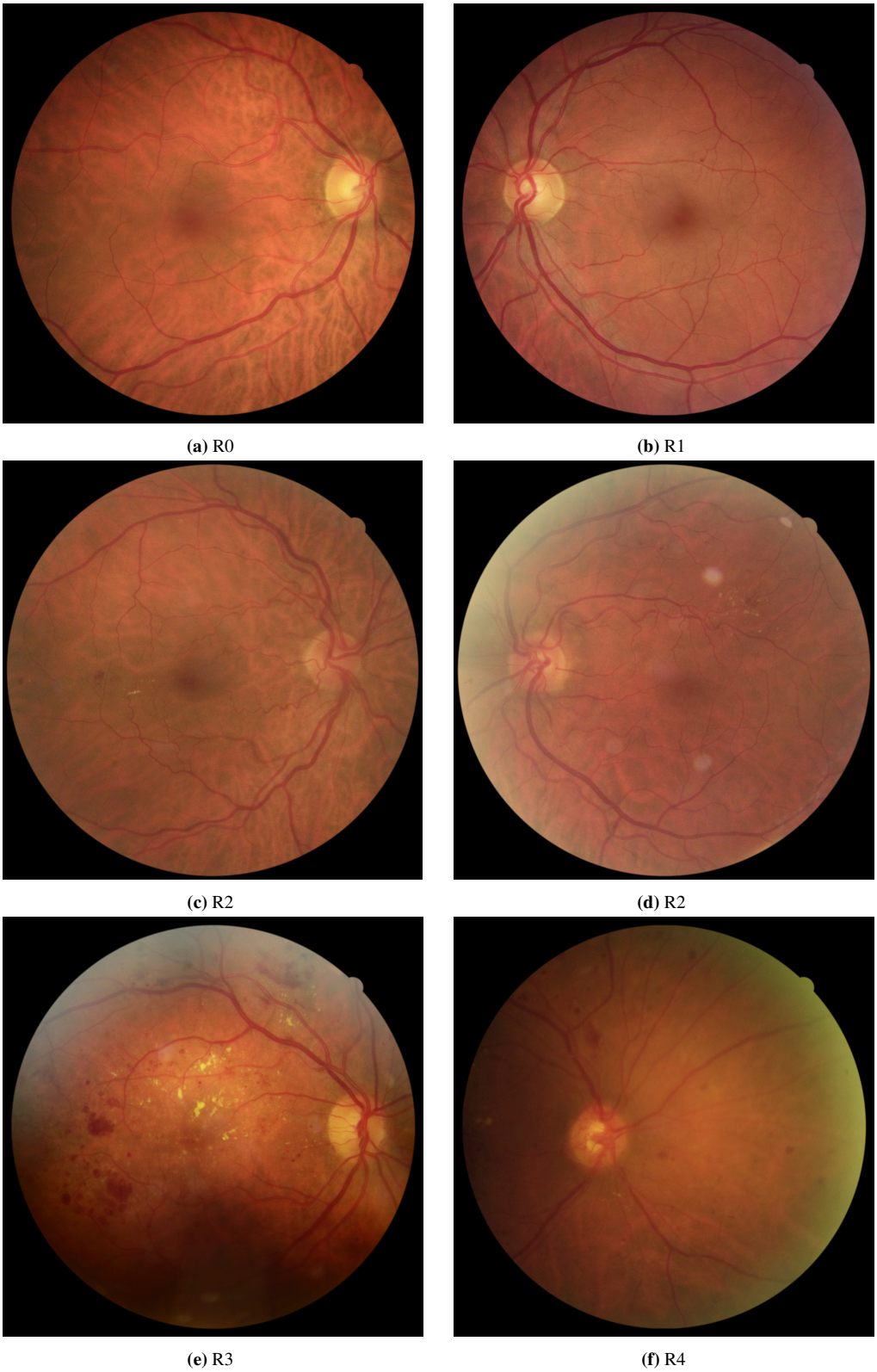


Figure 6.5: Agreement between annotators. Confidence stands for the number of annotators that agree with the most common label (majority voting).

Fig. 6.6 shows examples of images with full agreement between all annotators for the different DR grades. In contrast, Fig. 6.7 presents examples of images that originated high disagreement between annotators. The first two (Fig. 6.7a and 6.7b) were labeled as R2, R3, R3 and R4, meaning that three different labels were attributed to the image. Indeed, the differentiation between R2 and R3 in this case seems to be related with hemorrhage counting. The R4 label, by its turn, seems to result from the existence of structures that are/resemble neovessels. The examples from Fig. 6.7c and 6.7d were also attributed three different labels by four annotators: R1, R2 and R4. The differentiation between R1 and R2 is likely due to confusion of microaneurisms and hemorrhages. This shows that the DR grading task is indeed a complex task, even for retina specialists.

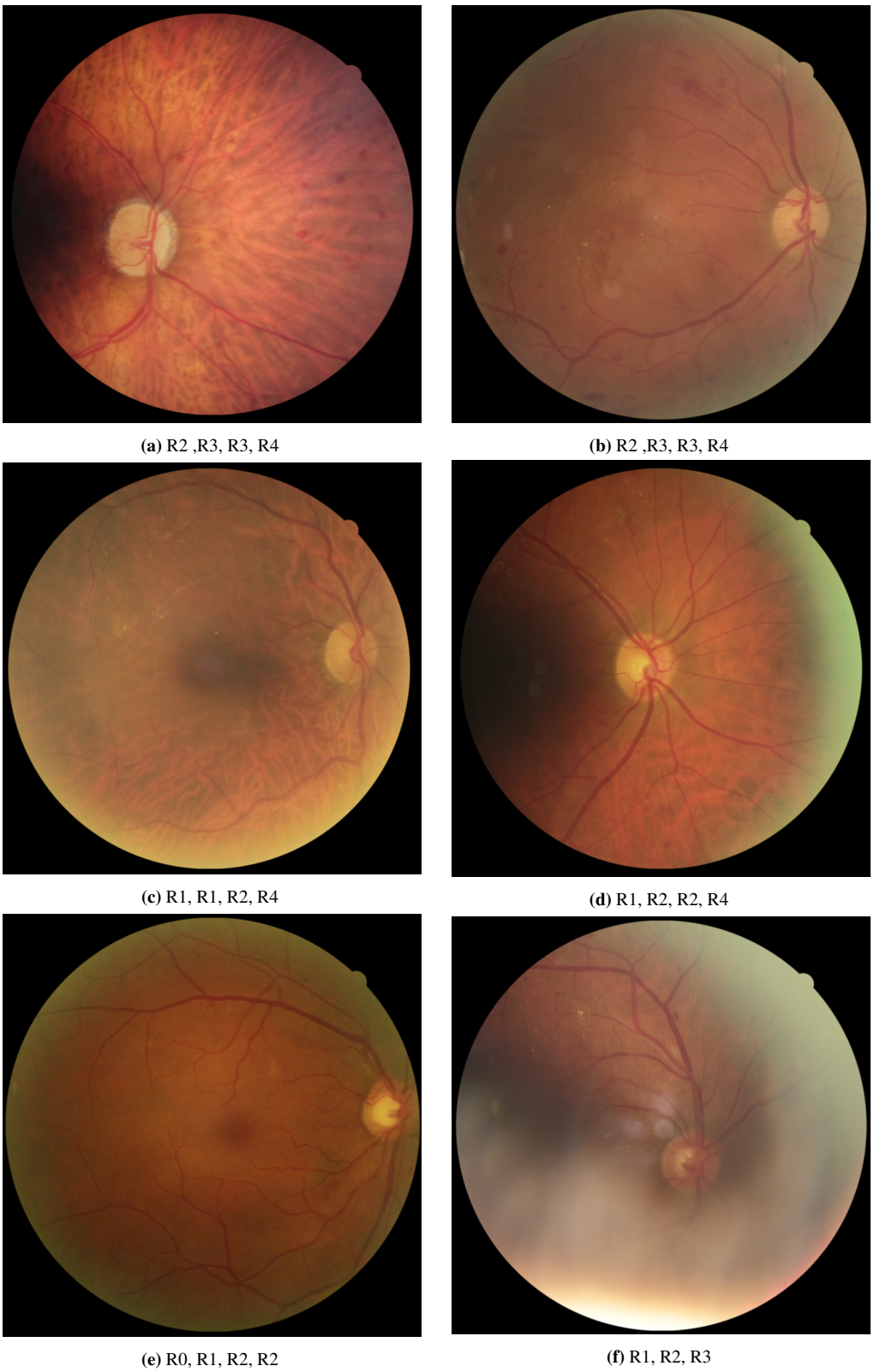
Fig. 6.7e was graded as R0, R1 and R2. Despite the grades being relatively close, one of the grades is R0 and one is R2, which suggests that an HEM/CWS was missed or identified as such without being a lesion. The image from Fig. 6.7f shows a curious example, in which the majority of the image is occluded. Still, the image was considered as gradable by the three annotators, and labeled with three different grades (R1, R2 and R3). Again, the R1 and R2 labels likely result from the MAs and HEMs confusion, but R3 must arrive from different count of HEMs or presence/absence of CWSs. Note, also, that when images are occluded they can only be graded based on the visible portion of the image, and thus the DR grade is *X* or higher, since it is possible that in the occluded portion could be present a lesion of a higher grade, which would be a sign of a severest grade. This also highlights the possible danger in consider solely the majority voting of annotations, even more when few annotators are available, to consider a final grade. These images are examples of cases which are not easy to classify as they lead to high disagreements, and thus should not be clearly attributed a grade based on majority voting with 100% confidence.

The annotation times were analysed in relation with the annotation disagreement, considering each different annotator as reference (Fig. 6.8). This disagreement is computed as the absolute difference between the two annotations, considering the grades $\in \{0,1,2,3,4\}$ corresponding to {R0, R1, R2, R3, R4}, respectively. Intuitively, one would expect that images that generated higher disagreements would be associated with higher annotation times. However, results did not show a clear relation between annotation times and disagreement in the annotations. However, for high/very high disagreements it appears that the annotation times tend to be higher for the majority of the annotators.



(a) R0 **(b)** R1 **(c)** R2 **(d)** R2 **(e)** R3 **(f)** R4

Figure 6.6: Examples of images with complete agreement between all four annotators. Below each image the grade attributed during the annotation process is presented.



(a) R2 ,R3, R3, R4

(b) R2 ,R3, R3, R4

(c) R1, R1, R2, R4

(d) R1, R2, R2, R4

(e) R0, R1, R2, R2

(f) R1, R2, R3

Figure 6.7: Examples of images with high disagreement between annotators. Below each image the different grades attributed during the annotation process are presented.

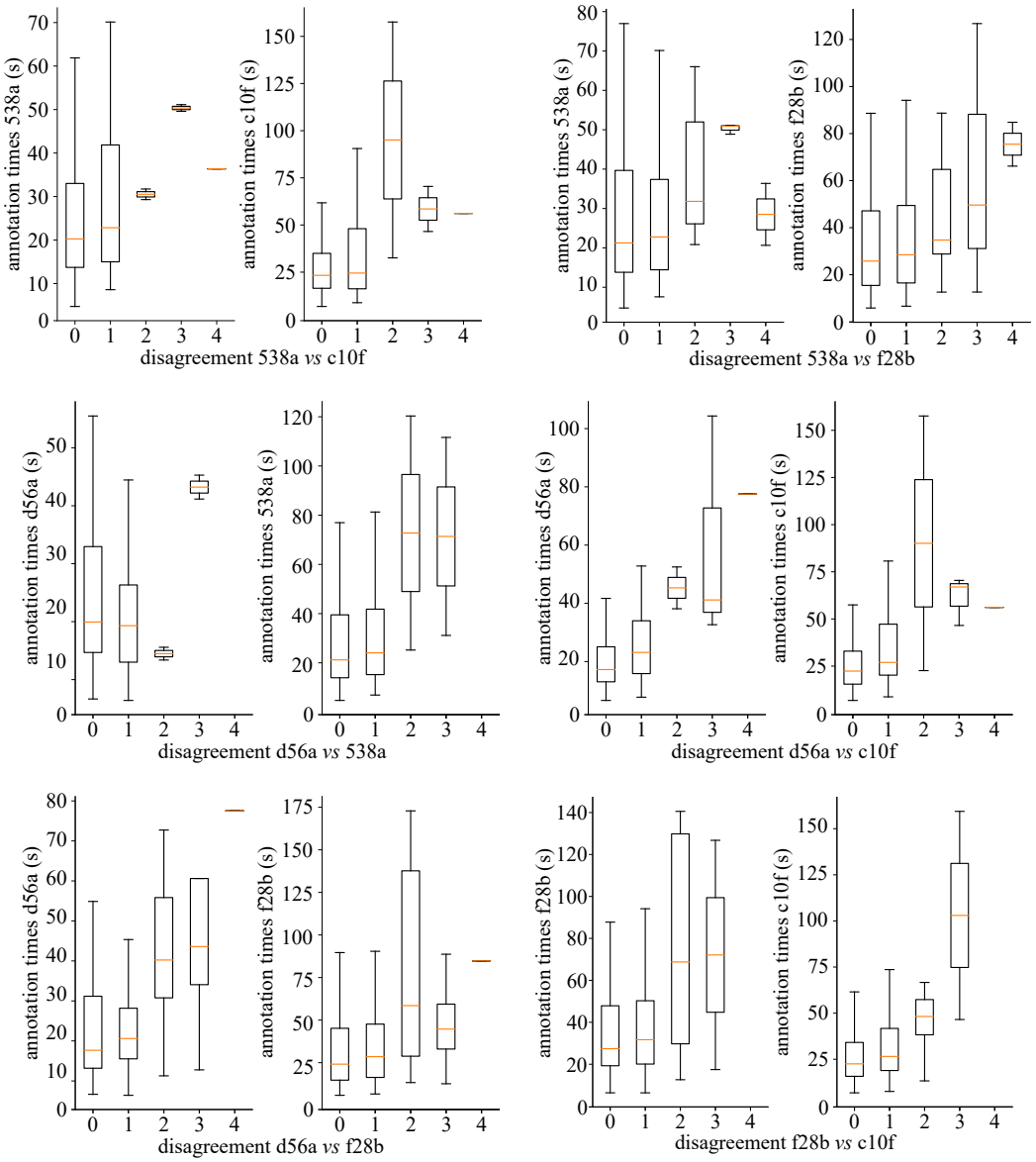
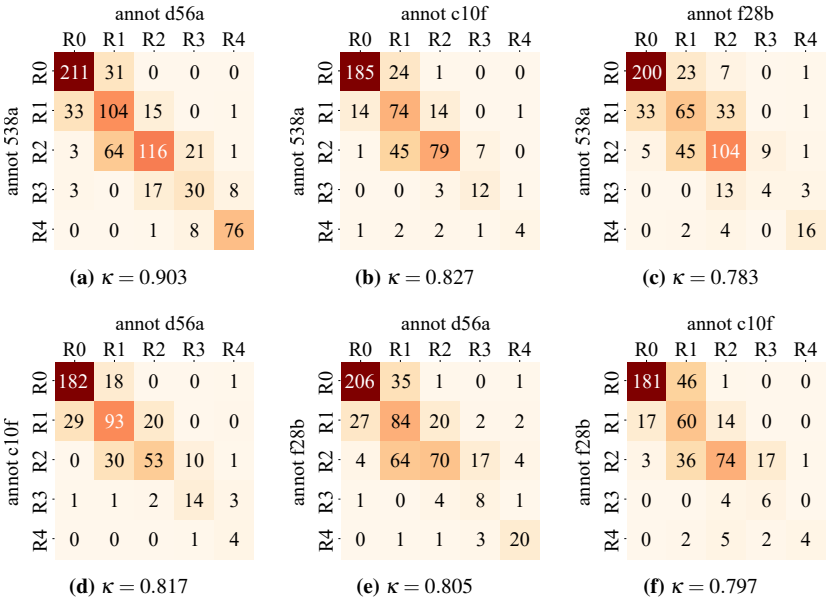


Figure 6.8: Annotation times *vs* annotation disagreement. The fliers are not shown in the plots to ease representation.

The DR performance of the annotators, comparing each pair, is shown in Fig. 6.9. Fig. 6.10 shows the normalized version of the confusion matrices, obtained by dividing each value of the original confusion matrix by the sum of the values from the corresponding row (total number of images from that grade in the dataset). These matrices show, overall, a diagonal tendency. The detection of R0 is a well achieved task across all annotators. For some annotators, however, one can see that the most confused classes appear to be the R2 and R3. Indeed, this distinction is based on the counting of HEMs or on detection of CWSs, and particularly the counting part can be affected by some degree of subjectivity. We can also see that for some annotators there is a significant confusion between R1 and R2, which may be due to confusion between MAs and HEMs, which is known to be a challenging task. Indeed, in Fig. 6.7 we showed examples of images with high disagreement in the annotations, which reflect this behaviour.



(a) $\kappa = 0.903$ **(b)** $\kappa = 0.827$ **(c)** $\kappa = 0.783$

(d) $\kappa = 0.817$ **(e)** $\kappa = 0.805$ **(f)** $\kappa = 0.797$

Figure 6.9: DR grading performance of each annotator comparing to each one of the other three.

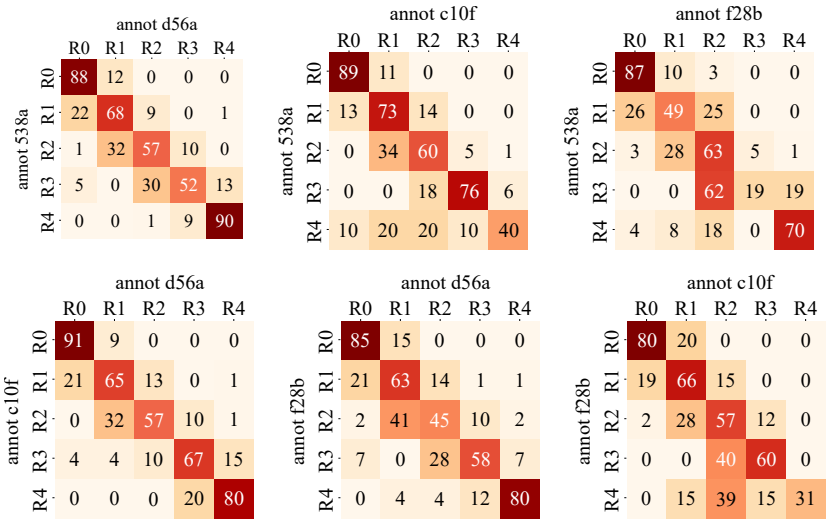


Figure 6.10: DR grading performance of each annotator comparing to each one of the other three. Matrices were normalized by dividing each value of the original confusion matrix by the sum of the values from the corresponding row (total number of images from that grade in the dataset), and the shown values correspond to percentages.

6.3.4 Automatic method and observers' analysis

The relation of the image quality annotation with the algorithm's prediction uncertainty is shown in Fig. 6.11. Uncertainty seems to be highly related with image quality, with higher uncertainties correlating with lower quality. The outlier cases (circles in the plots from Fig. 6.11) are expected to correspond to hard to diagnose cases, not directly due to poor image quality, since these are cases that the algorithm predicted with high uncertainty despite the quality label being partial/good.

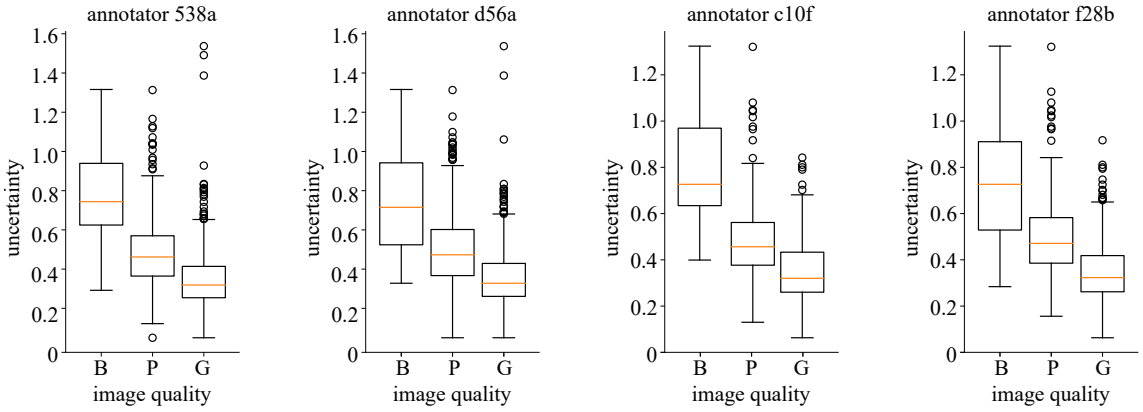
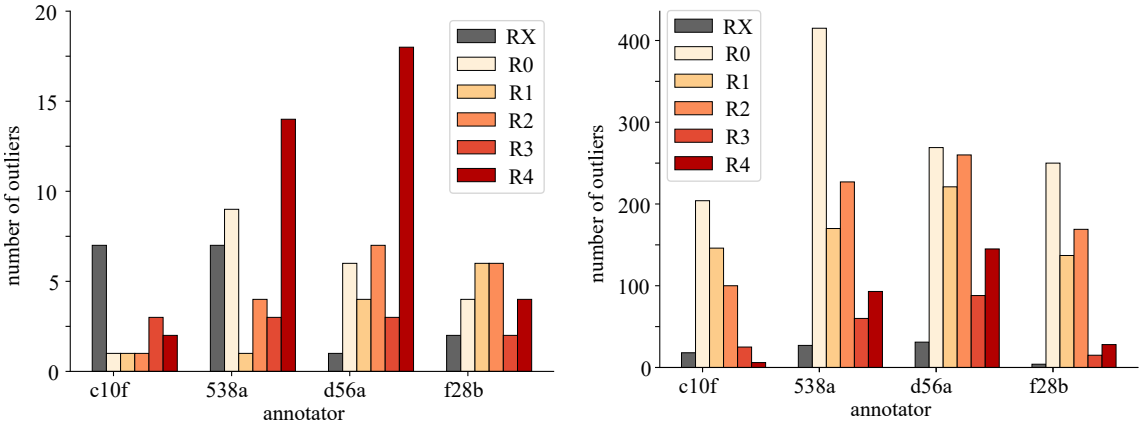


Figure 6.11: Uncertainty of the algorithm's prediction *vs* annotated image quality. B: bad, P: partial, G: good quality.



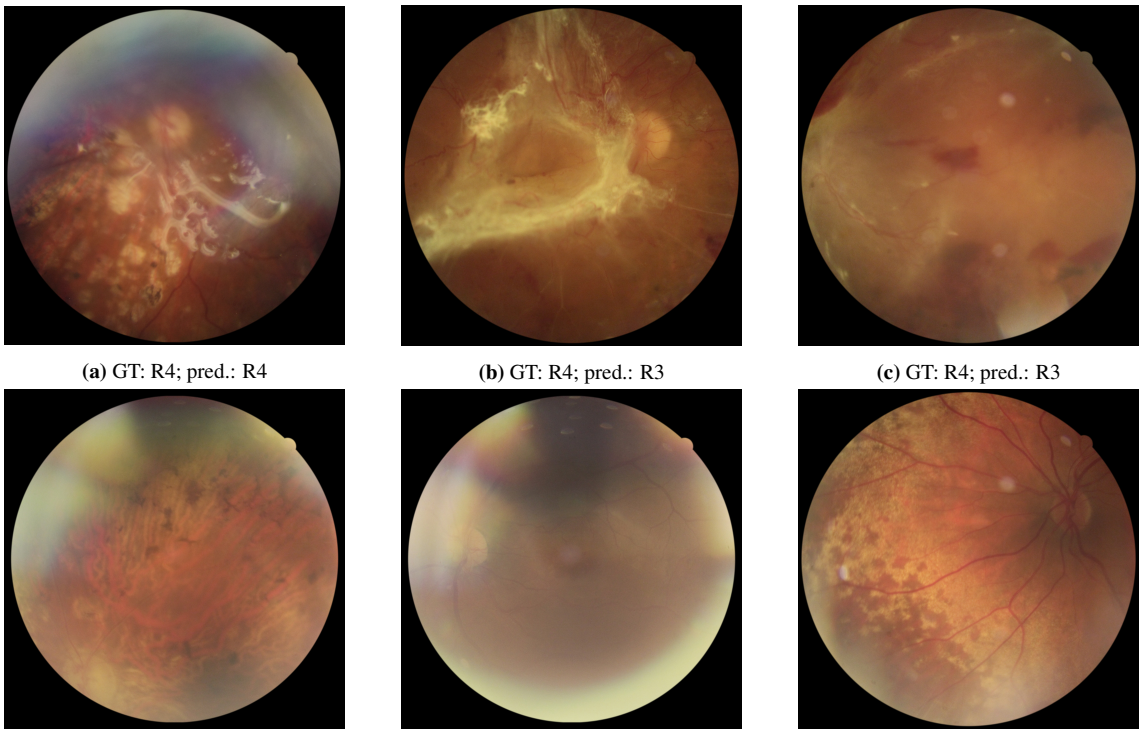
(a) Number of outliers per image grade.

(b) Number of annotated images of each image grade.

Figure 6.12: Outliers of uncertainty *vs* image quality plot (Fig. 6.11), distributed accordingly to the DR grade attributed by each annotator. RX: non gradable image.

Fig. 6.12a shows how these outliers distributed accordingly to the attributed DR grade, for each of the annotators. Note that the annotators *c10f* and *f28b*, as can be seen from Fig. 6.12b, annotated only a few R3 and R4 images, and thus the analysis for these two annotators is less confident than for the other two. The distribution of the outliers for the other two annotators (*d56a* and *538a*) suggests a very high prevalence of R4-labeled images among the outliers. This is in line with what we know about the algorithm, discussed in the past chapters, regarding the R4 detection. These R4-labeled images are images for which the method is not confident in the decision, likely due to lesions it cannot detect, such as pre-retinal fibrosis, and pre-retinal hemorrhages.

Additionally to these cases, we can also see that mainly for annotators *c10f* and *538a* a significant percentage of the outliers comes from RX images. These are images which the annotators considered as not possible to grade, even after considering them as having partial/good quality. These likely correspond to complex cases, in which the grade is hard to determine, and thus makes sense that the algorithm predicted an high uncertainty for these cases. Examples of outlier images from the annotator *538a* are shown in Fig. 6.13. Indeed, these cases appear to be complex ones, which are hard to grade (Fig.6.13d-f), or present R4 lesions the algorithm did not



(a) GT: R4; pred.: R4 **(b)** GT: R4; pred.: R3 **(c)** GT: R4; pred.: R3

(d) GT: RX; pred.: R2 **(e)** GT: R0; pred.: R2 **(f)** GT: R0; pred.: R1

Figure 6.13: Examples of outlier images from the uncertainty *vs* image quality plot for the annotator *538a* (Fig. 6.11).

learn (Fig. 6.13a-c). In the latter case, pre-retinal fibrosis, as shown in those images, are likely being interpreted as lack of image quality, as discussed in Chapter 4. In the case of Fig. 6.13e, the image does not seem to be gradable, despite the R0 labeling by the annotator, and should likely have been considered as of low quality.

Additionally, the analysis of the prediction uncertainty *vs* the annotator's disagreement was performed (Fig. 6.14). This disagreement is distance-based, differently from the one considered in Fig. 6.5, which was binary (agree/disagree). The disagreement here considered, for each pair of annotators, is the absolute distance between the two annotations (DR grades $\in \{0,1,2,3,4\}$), and thus can range from 0 to 4. Despite no linear relation, we can see that there is a tendency of higher uncertainties for higher disagreements.

Fig. 6.15 shows the confusion matrices of the automatic method comparing with each of the four annotators, and Table 6.3 the κ values for the annotators and for the algorithm *vs* the annotators. From this data, we see that the algorithm's performance is not far from that of the annotators. One should note that the network was trained on a completely independent dataset (Kaggle DR detection). As discussed before, the PDR detection is still the major drawback of the method. The evolution of the κ of the automatic method against uncertainty is shown in Fig. 6.16, as performed in Chapter 4 (Fig. 4.8). There is a tendency of lower κ values as the uncertainty threshold increases. The uncertainty appears to be a good indicative of the algorithm's performance.

The performance of the method accordingly to the image quality is shown in Fig. 6.17. For comparison, the performance of the annotators according to the quality is shown in Fig 6.18.

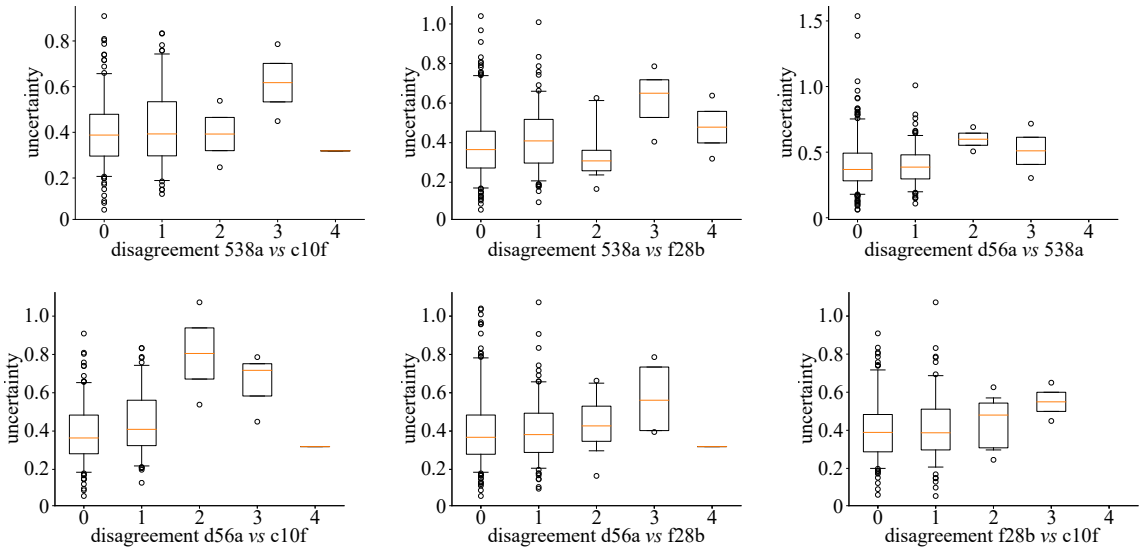


Figure 6.14: Uncertainty of the algorithm's prediction *vs* the annotator's disagreement in the diagnosis for each pair of annotators.

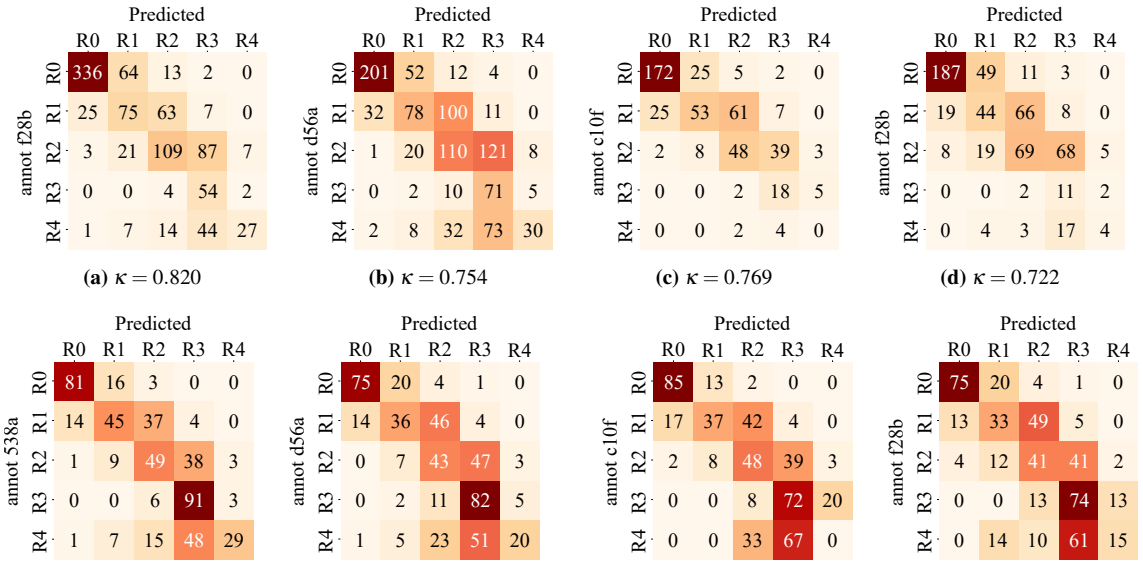


Figure 6.15: DR grading performance of the automatic method having as reference each of the annotators. Matrices on the top row are the original ones, and in the bottom row are the normalized versions (by dividing each value of the original confusion matrix by the sum of the values from the corresponding row; the values are percentages).

Table 6.3: DR grading performance of the annotators and the automatic method.

(a) Annotators.

Reference	Prediction	QWK	#images
538a	d56a	0,903	743
538a	c10f	0,827	470
538a	f28b	0,783	574
d56a	c10f	0,817	465
d56a	f28b	0,805	576
f28b	c10f	0,797	473
average		**0.822 ± 0.04**	

(b) Method and annotators.

Reference	Prediction	QWK	#images
538a	method	0,820	965
d56a		0,754	983
c10f		0,769	481
f28b		0,722	599
average		**0.77 ± 0.04**	

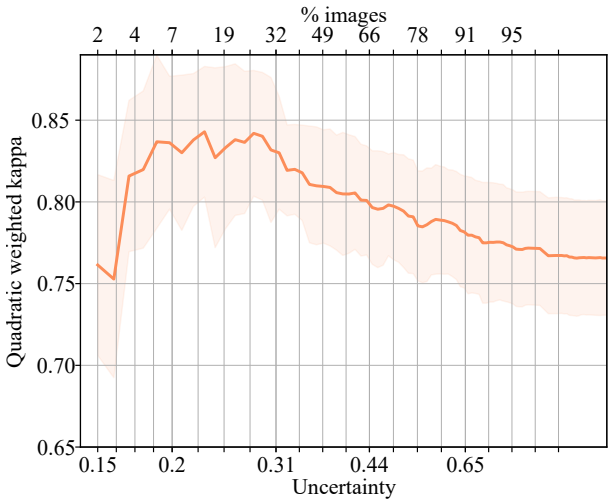


Figure 6.16: Quadratic weighted kappa *vs* uncertainty (*xx*-axis in logarithmic scale)

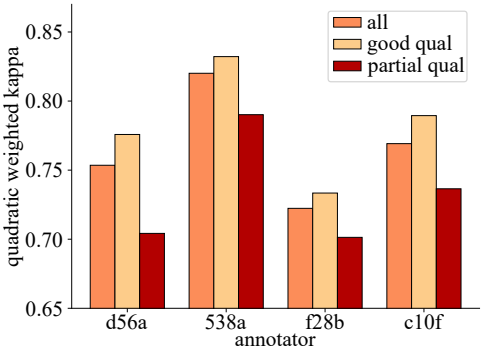


Figure 6.17: DR grading performance of the automatic method (quadratic weighted kappa score) having as reference each of the annotators accordingly to the image quality.

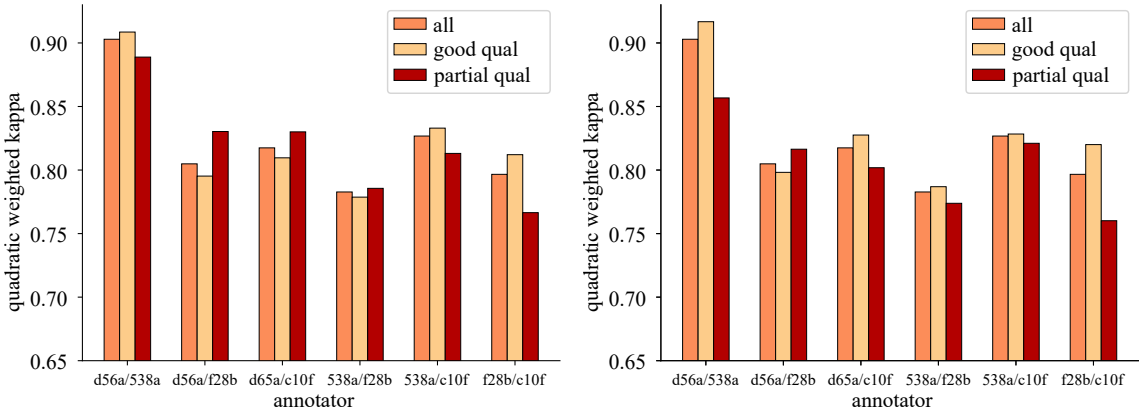
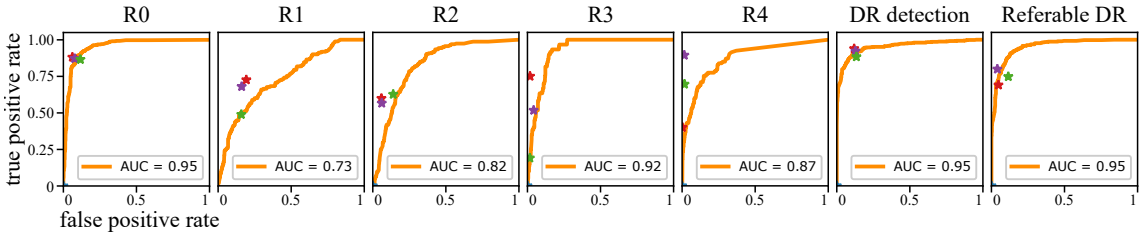
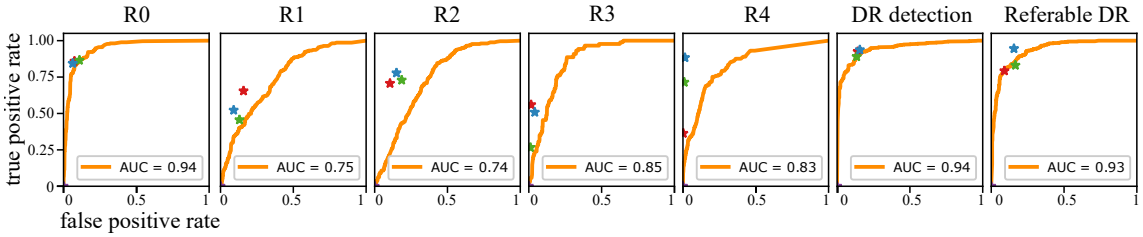


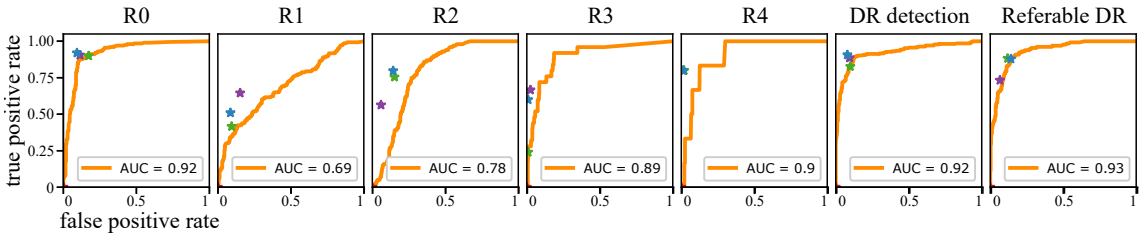
Figure 6.18: DR grading performance of the annotator (quadratic weighted kappa score), having as reference each one of the other annotators, accordingly to the image quality label.



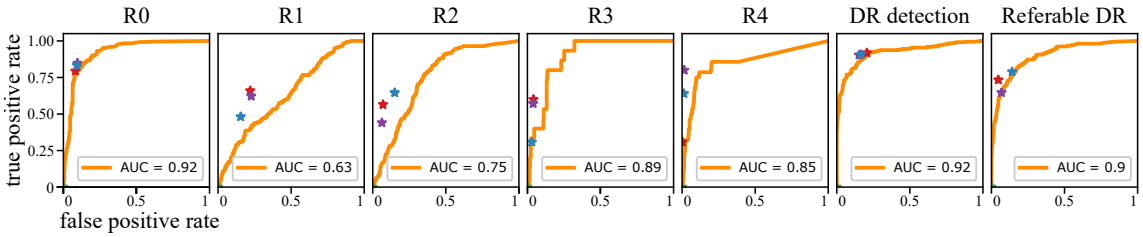
(a) Reference: annotator "538a".



(b) Reference: annotator "d56a".



(c) Reference: annotator "c10f".



(d) Reference: annotator "f28b".

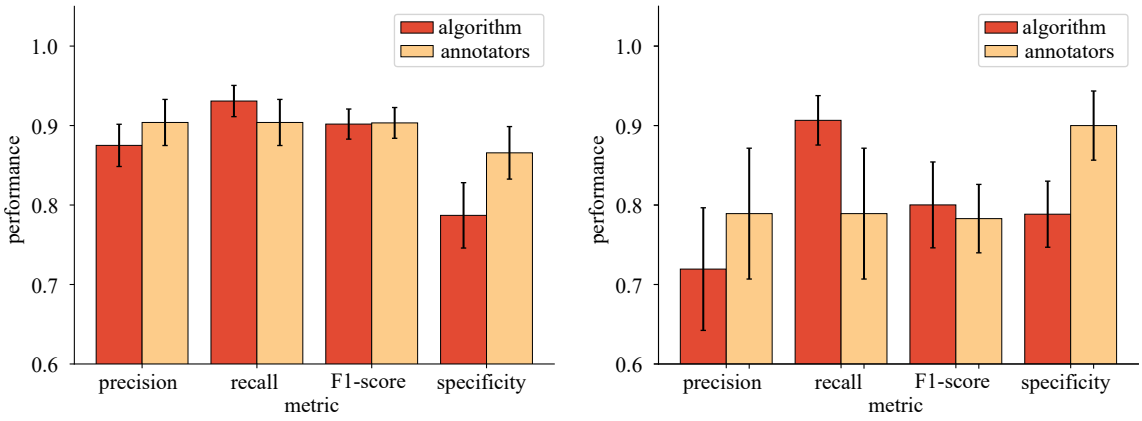
Figure 6.19: Per class performance of the algorithm and the annotators having as reference each of the annotators.

There is no consistent pattern/relation between the annotated image quality and the annotator's performance. One could expect that worse quality images could lead to a drop in performance, but that does not seem to be the case. When it comes to the algorithm, contrarily to what happens for the annotators, there is a tendency for better performance in good quality images than in partial quality ones, when evaluating against each annotator. This is in line with the previous findings, that showed that prediction uncertainty was highly correlated with image quality, and the uncertainty was a good indicator of the algorithm's performance.

The per class detection performance of the algorithm and of the annotators having as reference each of the annotators is shown in Fig. 6.19. ROCs are plotted for each class by varying the threshold of the network's output probability for that class. The isolated points correspond to the other three annotators. Additionally to the per grade detection performance we also show the curves for DR detection and referable DR detection (*i.e.*, detection of R2 or higher). The AUCs of these

Table 6.4: Area under the curve (AUCs) for each detection task (ROCs from Fig. 6.19).

Reference	Prediction	AUC							#images
		R0	**R1**	**R2**	**R3**	**R4**	**detect.**	**referable**	
538a	method	0.95	0.73	0.82	0.92	0.87	0.95	0.95	965
d56a		0.84	0.75	0.74	0.85	0.83	0.94	0.93	983
c10f		0.92	0.69	0.78	0.89	0.9	0.92	0.93	481
f28b		0.92	0.63	0.75	0.89	0.85	0.92	0.90	599
average		**0.91**	**0.70**	**0.78**	**0.89**	**0.86**	**0.93**	**0.93**	
std		0.05	0.05	0.04	0.03	0.02	0.02	0.02	



(a) DR detection.

(b) Referable DR detection.

Figure 6.20: DR and referable DR performance of the algorithm and the annotators in terms of precision, recall, specificity and F1-score. The shown values are the average (along with the standard deviation) of the metrics computed against each of the (other) annotators.

ROCs are shown in Table. 6.4. The DR and referable DR detection performance of the algorithm is very good, with an average AUC of 0.93 in these tasks (across all annotators). The worst AUC is for the R1 class, followed by R2, which is likely related with the difficulty in distinguishing microaneurisms and hemorrhages. These are also the classes for which the algorithm's curves are further bellow the annotators' points.

The DR and referable DR detection performance are analysed in Fig. 6.20 in terms of precision, recall, F1-score and specificity. The values are the average of the metrics computed against each of the annotators. DR and referable DR detection performance is similar between annotators and the automatic method. Indeed, the F1-score (harmonic mean between precision and recall) values of DR|GRADUATE are similar or better comparing to the specialists. For the referable DR detection the recall of the algorithm is considerably better than that of the annotators. This comes at the expense of a reduction in specificity, but considering a screening context, in which it is of special importance to detect all possible positive cases, DR|GRADUATE could prove to be a very useful tool.

6.4 A note on DR grading

This thesis has focused on the development of automatic methods for the analysis of eye fundus images, particularly for the grading of DR based on a single fundus photography. However, and as referred in Section 2.4, the gold standard for visual documentation of DR is the ETDRS 7 standard field stereoscopic photographs since 1991 (ETD, 1991). Using these 7 standard field photographs allow higher sensitivity and specificity comparing to single-field fundus photographs interpreted by trained readers (Ghasemi Falavarjani *et al.*, 2017), thus being more reliable for the assessment of DR. However, using these montages is not as convenient as single photography for patients and photographers since they require more acquisition time, more skilled technicians and pharmacological pupil dilation (Ghasemi Falavarjani *et al.*, 2017). A possible alternative is to assess a single panoramic view of the eye via a non-mydriatic shot. For instance, Ultrawide-field (UWF) is an eye fundus photograph that has a 100 degree or greater field (Ghasemi Falavarjani *et al.*, 2017). Whereas the 7 field standard images cover only 30% of the retina, UWF can raise up this value up to approximately 80%. Studies have shown that 200 degree UWF images detect more DR lesions than using routine 30 degree photography, and that these lead to a more severe DR grade in 10% of the eyes (Sun *et al.*, 2018). When comparing UWF imaging with the conventional ETDRS fields for detection of PDR, current data suggests that UWF allowed to detect additional and more extensive PDR pathologies, with these likely carrying prognostic value (Liu and Arevalo, 2019). There is still discordance, however, on the correlation between peripheral DR lesions and management strategies to treat DR (Ghasemi Falavarjani *et al.*, 2017).

Additionally to UWF imaging, optical coherence tomography angiography (OCT-A), which is an emerging advanced retinal imaging technique, can be useful for DR diagnosis. OCT-A allows to visualize the retinal vessels in 3D without requiring a contrasting agent (Sun *et al.*, 2018). Studies have shown the potential of OCT-A in distinguishing IRMAs from NVs (Arya *et al.*, 2020), which is a known issue when using fundus photography, particularly when it comes to the early NVs (Sun *et al.*, 2018). This differentiation is key for distinguishing severe NPDR from PDR, and thus OCT-A may prove to be a useful tool to diagnose the severest cases of DR.

6.5 Concluding remarks

The inter-observer analysis has shown that, whilst specialists tend to agree or are close in their DR diagnosis, there are cases of high disagreement, which highlights the complexity of the DR grading task and the inherent diagnosis subjectivity in difficult cases. Indeed, it is very common that at least one out of two to four annotators differs in their diagnosis.

The proposed automatic method showed a performance overall comparable with the medical doctors. It has particularly good performance for DR detection and referable DR detection, thus showing its potential as a screening tool, as it allows to discard *a priori* cases without or with mild disease and refer the most severe cases to specialists. Indeed, the recall of the method, particularly for the referable DR detection, was higher than that of the annotators, suggesting it may capture

more diseased cases than humans, and the F1-score of the method was as good or better comparing to the annotators. However, and as discussed in previous chapters, it misses several R4 cases.

The uncertainty estimation appears to be a good indicative of how much one should trust the algorithms' decision, and is closely related with image quality. This constitutes a very valuable tool in the real-world practice since it will provide specialists a measure of quality of the automatic diagnosis and make the decision process more efficient and trustworthy.

Indeed, these results are encouraging since the network was trained on a completely different dataset than the one used for this validation, and has showed human-like performance on a independent set of data from a screening program. In sum, the method's DR prediction together with the uncertainty estimation can be valuable in the clinic, with great potential to overcome limitations of current techniques which prevent them from entering real-world applications.

Chapter 7

Conclusions

Diabetic retinopathy (DR) is a complication of diabetes and is one of the leading causes of blindness worldwide. The majority of visual loss cases can be prevented with early detection and adequate treatment. DR screening programs have been developed across the world, in which DR detection and DR severity grading are performed, based on the analysis of retinal lesions visible in color eye fundus images. Due to the complexity of this task and high workload of the specialists, computer-aided diagnosis systems are desirable for reducing their burden and to reduce diagnosis' subjectivity. However, the complexity of the grading task and the wide variety of lesions' characteristics makes the development of these systems a non-trivial task. Deep learning approaches have shown promising performance as computer-aided diagnosis (CAD) systems, but their black-box behaviour hinders clinical application.

In Chapter 3 we presented UOLO, a novel framework for the simultaneous detection and segmentation of structures of interest in medical images. We validated UOLO on the task of simultaneous optic disc (OD) detection, fovea detection, and OD segmentation from retinal images, which are of interest in the diagnosis of several eye diseases. UOLO consists of an object segmentation module based on the U-Net which intermediate abstract representations are processed and used as input for object detection. The resulting system is optimized simultaneously for detecting a class of objects and segmenting an optionally different class of structures. This network can be trained with relatively few images with segmentation ground truth and still maintain a high performance. UOLO is also robust to inter-dataset settings, thus showing great potential for applications in the medical image analysis field.

Chapter 4 focused on the main contribution of this thesis: DR|GRADUATE, a novel deep learning-based DR grading CAD system that supports its decision by providing a medically interpretable explanation and an estimation of how uncertain that prediction is, allowing the ophthalmologist to measure how much that decision should be trusted. Indeed, deep learning-based techniques tend to outperform classical feature engineering approaches, since they are capable of learning and relating relevant features in a complex manner without requiring field knowledge. However, the large number of parameters to train makes the success of these models highly data reliant. Also, as the recovery and interpretation of the learned features is non-trivial, this greatly

hinders the application of these methods for computer-aided diagnosis in the medical centers, as it is hard to trust a second opinion that may appear as unfounded. Additionally, most of the deep learning solutions produce deterministic outputs without quantifying the uncertainty in the predictions. The fact that a specialist knows how uncertain the model is about a prediction allows him to establish a priority list, highlighting the cases that are more complex and thus would require a closer examination or second opinion.

DR|GRADUATE takes this into account, and was designed considering the ordinal nature of the DR grading problem. A novel Gaussian-sampling approach built upon a Multiple Instance Learning framework allow DR|GRADUATE to infer an image grade associated with an explanation map and a prediction uncertainty while being trained only with image-wise labels. The classification of the R4 grade was the least accomplished task, suggesting the training did not correctly capture the proliferative DR (PDR) signs, most likely due to the type of annotation in the Kaggle DR detection training dataset, with images that underwent photocoagulation treatment and show laser marks appearing to be labeled as PDR even when not presenting the typical signs of this grade.

We show that high performance occurs for images with low prediction uncertainty, thus indicating that this uncertainty is a valid measure of the predictions' quality. Further, bad quality images are generally associated with higher uncertainties, showing that images not suitable for diagnosis indeed lead to less trustworthy predictions. The attention maps generally highlight regions of interest for diagnosis. Comparing with the state-of-the-art, DR|GRADUATE has the advantage of producing both an uncertainty estimation associated with the prediction and an explanation map highlighting the most relevant regions for the classification, without compromising the DR grading performance.

We also proposed a data augmentation scheme to compensate for the lack of PDR cases in DR-labeled datasets (Chapter 5). PDR is an advanced diabetic retinopathy stage, characterized by neovascularization, which leads to ocular complications and severe vision loss. However, the available DR-labeled retinal image datasets have a small representation of images of the severest DR grades, and the criteria for labelling these images in the publicly available datasets is not always clear, with some photo-coagulation treatment and laser marks being considered as PDR signs. We proposed a heuristic-based data augmentation scheme based on the synthesis of neovessel (NV)-like structures that compensates for this lack of PDR cases in DR datasets. The proposed algorithm relies on the general knowledge of common location and shape of these structures. NVs are synthesized and introduced in pre-existent retinal images which can then be used for enlarging deep neural networks' training sets. The data augmentation scheme was tested on multiple datasets, and allows to improve the DR|GRADUATE's capacity to detect NVs.

Finally, in Chapter 6 we presented a scheme for estimating the DR grade of an image based on an ensemble of the methodologies described on the past chapters. We also performed a thorough analysis on the private SCREEN-DR dataset, both of the automatic algorithm and of the retinal specialists involved in its annotation. The inter-observer analysis has shown that, whilst doctors tend to agree or are close in their DR diagnosis, there are cases of high disagreement, which

highlights the complexity of the DR grading task and the inherent diagnosis subjectivity in difficult cases. The proposed method showed a performance overall comparable with the medical doctors, with a particularly good performance for DR detection and referable DR detection, thus showing its potential as a screening tool, as it allows to discard *a priori* cases without or with mild disease and refer the most severe cases to specialists. These results are encouraging considering that the network was trained on a given dataset, and has showed human-like performance on a screening program's independent dataset.

As future research, the exploitation of the data augmentation method for improving PDR detection in terms of shape variety and realism of the structures could be useful. Additionally, despite NVs being the most common sign of PDR, the other manifestations, *i.e.*, pre-retinal hemorrhages and pre-retinal fibrosis, should also be explored for augmenting the training datasets. These are not definable in terms of shape as NVs, and thus a generation approach as the one followed for the NVs would not be suitable. Other strategies, such as those based on generative adversarial networks (GANs), could be explored. Also, besides augmenting training datasets for improving the variability of the existent cases, a good construction of the datasets is paramount for the success of deep learning-based techniques. Large enough datasets and with good quality annotations are key for training deep learning systems. For instance, DR grades could result from the adjudication of the decision from different retinal specialists to obtain a consensus labeling, allowing for a more robust labeling while still encoding the difficulty of the image as function of the agreement between specialists. Thus, involving multiple annotators and performing adjudication sessions between them could be a path to pursuit. Another interesting improvement in the annotation could be to enlarge the type of information provided by the specialists. Besides the DR grade, other information could be provided, such as the self-reported confidence of the annotator in that diagnosis. This could be used for training, for instance by weighting the samples accordingly to their grading confidence. Complementarily to the development of these techniques, the assessment of the influence of these systems in the decision process in real-world practice should also be explored.

Studies in multiple areas have shown the potential of using multimodal data for diagnosis of different diseases, which could also be explored in DR grading. This data could include images from different imaging modalities or image and textual information. Please also note that the studies analysed and the methods proposed in this thesis are focused on the diagnosis of a single fundus images per eye. This is indeed the common approach found in the literature, which makes sense since the vast majority of available image datasets for developing the grading algorithms present labeling for a single fundus image and not of multiple fields. Nevertheless, it is important to keep in mind the recent advances in retinal imaging and how these allow to improve current diagnosis standards and thus allow better outcomes for the patients. Future research should thus probably focus on the exploitation of techniques such as ultra-wide field (UWF) and optical coherence tomography angiography (OCT-A) imaging for improving DR diagnosis.

This thesis focused on the development the methodologies for a system that can be more easily integrated in real-world practice, that produces outcomes that can be more easily understood and possibly accepted by the health professionals. The system's decisions are non deterministic, with

the association of a confidence and a visual explanation to each diagnosis, and its performance was validated across multiple datasets and types of data to ensure good generalization capability. At test time, *i.e.* where the performance of the system is not known, the uncertainty can be a measure of how much a specialist can trust the algorithm's decision, allowing to rank the predictions and select which images would benefit most from a second evaluation. Further, the explanation map produced by our algorithm should be easily interpretable by the retinal specialists, and thus constitutes a valuable tool for mitigating the black-box behaviour associated with deep neural networks. It is the author's hope that the proposed system can ease the DR diagnosis process in the clinic, functioning as a tool to ease the healthcarers work and improve DR diagnosis' quality, and consequently the patient's lives.

Chapter 8

List of publications

International journals

- **Araújo, T.**, Aresta, G., Mendonça, L., Penas, S., Maia, C., *et al.* DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715, 7 2020c. doi: 10.1016/j.media.2020.101715
- **Araújo, T.**, Aresta, G., Mendonça, L., Penas, S., Maia, C., *et al.* Data Augmentation for Improving Proliferative Diabetic Retinopathy Detection in Eye Fundus Images. *IEEE Access*, 8:182462–182474, 2020b. doi: 10.1109/ACCESS.2020.3028960
- Aresta, G., **Araújo, T.**, Kwok, S., Chennamsetty, S. S., Safwan, M., *et al.* BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019c. doi: 10.1016/j.media.2019.05.010 (first two authors contributed equally)
- **Araújo, T.**, Mendonça, A. M., and Campilho, A. Parametric model fitting-based approach for retinal blood vessel caliber estimation in eye fundus images. *PLoS ONE*, 13(4):1–27, 2018b. doi: 10.1371/journal.pone.0194702
- **Araújo, T.**, Aresta, G., Castro, E., Rouco, J., Aguiar, P., *et al.* Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS ONE*, 12(6):e0177544, 2017a. doi: 10.1371/journal.pone.0177544 (first three authors contributed equally)

- Challenge's participation

- Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., *et al.* IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge. *Medical Image Analysis*, 59:101561, 1 2020. doi: 10.1016/j.media.2019.101561
- Al Hajj, H., Lamard, M., Conze, P. H., Roychowdhury, S., Hu, X., *et al.* CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. *Medical Image Analysis*, 52:24–41, 2019. doi: 10.1016/j.media.2018.11.008

Conferences proceedings

- **Araújo, T.**, Aresta, G., Galdran, A., Costa, P., Mendonça, A. M., *et al.* UOLO - Automatic object detection and segmentation in biomedical images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11045 LNCS:165–173, 2018a. doi: 10.1007/978-3-030-00889-5_19
- Wanderley, D. S., **Araújo, T.**, Carvalho, C. B., Maia, C., Penas, S., *et al.* Analysis of the performance of specialists and an automatic algorithm in retinal image quality assessment. In *IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, Lisboa, 2019. IEEE. doi: 10.1109/ENBENG.2019.8692506 (first three authors contributed equally)
- Costa, P., **Araújo, T.**, Aresta, G., Galdran, A., Mendonça, A. M., *et al.* EyeWeS: Weakly Supervised Pre-Trained Convolutional Neural Networks for Diabetic Retinopathy Detection. In *International Conference on Machine Vision Applications (MVA)*, pages 5–15, Tokyo, 2019. doi: 10.23919/MVA.2019.8757991
- **Araújo, T.**, Aresta, G., Almada-Lobo, B., Mendonça, A. M., and Campilho, A. Improving Convolutional Neural Network Design via Variable Neighborhood Search. In *ICIAR 2017 - International Conference on Image Analysis and Recognition LNCS 10317*, pages 371–379. 2017b. doi: 10.1007/978-3-319-59876-5_41 (first two authors contributed equally)
- Mendonça, A. M., Melo, T., **Araújo, T.**, and Campilho, A. Optic Disc and Fovea Detection in Color Eye Fundus Images. In *Image Analysis and Recognition*, pages 332–343, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-50516-5_29
- Galdran, A., Costa, P., Bria, A., **Araújo, T.**, Mendonça, A. M., *et al.* A No-Reference Quality Metric for Retinal Vessel Tree Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 82–90, Cham, 2018b. Springer International Publishing. doi: 978-3-030-00928-1_10
- Galdran, A., **Araújo, T.**, Mendonça, A. M., and Campilho, A. Retinal Image Quality Assessment by Mean-Subtracted Contrast-Normalized Coefficients. In *VipIMAGE 2017*, pages 844–853, Cham, 2018a. Springer International Publishing. doi: 10.1007/978-3-319-68195-5_92

Abstracts

- **Araújo, T.**, Maia, C., Penas, S., Mendonça, L., Carneiro, , *et al.* Explainable and uncertainty-aware system for diabetic retinopathy grading. In *20th EURETINA, European Society of Retina Specialists*, 2020a
- **Araújo, T.**, Wanderley, D., Carvalho, C., Costa, P., Maia, C., *et al.* Performance analysis of specialists and automatic methods in diabetic retinopathy screening. In *19th EURETINA, European Society of Retina Specialists*, Paris, 2019a
- Aresta, G., **Araújo, T.**, Campilho, A., Eloy, C., Polónia, A., *et al.* BACH: Grand challenge on breast cancer histology images. In *Proceedings of Machine Learning Research - MIDL 2019*, London, 2019a

- **Araújo, T.**, Aresta, G., Galdran, A., Costa, P., Mendonça, A. M., *et al.* UOLO: a framework for object detection and segmentation in medical images. In *Portuguese Conference on Pattern Recognition (RECPAD)*, Porto, 2019b
- Aresta, G., **Araújo, T.**, Campilho, A., Eloy, C., Polónia, A., *et al.* BACH: Grand challenge on breast cancer histology images. In *Encontro Nacional de Jovens Investigadores em Oncologia*, Porto, 2019b

Appendix A

Methods for diabetic retinopathy diagnosis

Table A.1: Summary of the main methods for diabetic retinopathy (DR) grading, referable DR detection and DR detection. The type of classification performed is presented, as well as the main methodology applied. The datasets used in the works and the main reported results are referred. sens: sensitivity; spec: specificity; acc: accuracy; AUC: area under the curve; QWK: quadratic-weighted Cohen's kappa; ACA: average classification accuracy

Article	**Type**	**Main Method**	**Datasets**	**Results**
Krause *et al.* (2018)	DR grading (5 classes)	Deep learning: ensembles of CNNs	- 1813 images from EyePACs; - training: 1662646 images from EyePACS (private)	QWK: 0.84
de la Torre *et al.* (2018)	DR grading (5 classes)	Deep learning; QWK-based loss	- Kaggle test set; - training: Kaggle	QWK: 0.74
Zhao *et al.* (2019)	DR grading (5 classes)	Deep learning: "BIRA-NET" (attention model for feature extraction and bilinear model for fine-grained classification); distance-based loss	- Kaggle (1560 images for test, rest for training)	ACA: 0.543
Galdran *et al.* (2020b)	DR grading (5 classes)	Deep learning; loss accounting for class distance and inter-observer disagreement knowledge	- Kaggle test; - Messidor-2 - training: Kaggle	- Kaggle: QWK: 0.787, ACA: 0.546 - Messidor-2: QWK: 0.798, ACA: 0.634
Gonzalez-Gonzalo *et al.* (2020)	DR grading (5 classes)	Deep learning; iterative inpainting technique for explanation map generation	- Kaggle test set; ; - training: Kaggle	QWK: 0.77-0.80
Costa *et al.* (2019)	DR detection	Deep learning: Multiple-Instance Learning framework for explanation map generation	- Messidor (also for training);	- AUC: 0.958

Gulshan *et al.* (2016)	Referable DR detection	Deep learning: ensembles of CNNs	- 9963 images from EyePACs; - Messidor-2 (1748 images, 874 patients); - training: 128175 images from EyePACS in US + 3 eye hospitals in India	- EyePACs: AUC: 0.991, sens: 90.3%, spec: 98.1%; sens: 97.5%, spec: 93.4%; - Messidor-2: AUC: 0.990, sens: 87%, spec: 98.5%; sens: 96.1%, spec: 93.9%
Abràmoff *et al.* (2016)	Referable DR detection	Deep learning: individual CNNs to detect each lesion type	- Messidor-2; - training: images from the EyeCheck project and the University of Iowa	AUC: 0.980, sens: 96.8%, spec: 87.0%
Quellec *et al.* (2017)	Referable DR detection	Deep learning: Kaggle-adapted architecture; (new backpropagation method; heatmaps for visualization)	- Kaggle (also for training); - e-ophtha	- Kaggle: AUC: 0.954; - e-ophtha: AUC: 0.949
Takahashi *et al.* (2017)	DR grading in 3 classes (modified Davis scale): no, simple, preproliferative and proliferative DR	Deep learning, using 4 fundus images per eye	- 9443 images for training, 496 for testing	acc.: 81%; prevalence and bias-adjusted Fleiss' kappa (PABAK): 0.64
Roy *et al.* (2017)	DR grading (5 classes)	Deep learning (CNNs) combined with dictionary-based approaches for feature extraction + random forest classifier	- Kaggle (also for training)	QWK: 0.86
Gargeya and Leng (2017)	DR detection	Deep learning: deep residual nets + metadata (feature extraction) + decision tree classifier	- Messidor-2; - e-ophtha; - 75137 images from the EyePACS (also used for training)	- Messidor-2: AUC: 0.94, sens: 93%, spec: 87%; - e-ophtha: AUC: 0.95, sens: 90%, spec: 94%; - EyePACs: AUC: 0.97, sens: 94%, spec: 98%
Fadzil *et al.* (2011)	DR grading in 3 NPDR classes	Gaussian Bayes classifier based on the area of the foveal avascular zone (FAZ)	- FINDeRD (315 images)	(in average): - grading: acc: 97.46%; - detection: sens: 90.81%, spec: 98.29%
Saleh and Eswaran (2012)	DR grading in 3 NPDR classes	Rule-based grading (based on MAs and HEMs)	- 98 images	- grading: acc: 90.8%; - detection: sens: 89.47%, spec: 95.65%
Gupta *et al.* (2017)	PDR detection	Local texture and vesselness-based features + random forests for NV characterization; PDR is present if the vascularity level in the image is > threshold	public data (only some images) and private data (total: 779 images): - Messidor; - STARE; - HRF	(3 datasets): AUC: 0.96, sens: 83.3%, spec: 96.1%
Akram *et al.* (2014)	DR grading in 3 NPDR classes	Multivariate m-Medoids + Gaussian Mixture Model hybrid classifier for lesion classification; grading based on lesion type and count (MA, HEM, EX)	- DRIVE; - STARE; - DIARETDB; - Messidor	(all datasets): - grading: acc: 98.52%; - detection: sens: 99.17%, spec: 97.07%
Acharya *et al.* (2012)	DR grading: normal, NPDR, PDR	Texture features + SVM classifier (integrated DR index)	- 238 images	- grading: acc: 85.2%; - detection: AUC: 0.972, sens: 98.9%, spec: 89.5%

Roychowdhury *et al.* (2014)	DR grading in 3 NPDR classes	Two-step hierarchical lesion classification (1^{st}: bright lesions, red lesions, discard non lesions; 2^{nd}: divide into CWS, EX, MA, HEM); grading based on lesion count	- Messidor; - training: DIARETDB1	detection: AUC: 0.904, sens: 100%, spec: 53.16%
Mookiah *et al.* (2013)	DR grading: normal, NPDR, PDR	PNN/Decision Trees/SVM trained to grade DR (features extracted from EX and vessel segmentations; eg: bifurcation points count)	- 156 images	(PNN): - grading: acc: 96.15%; - detection: sens: 96.27%, spec: 96.08%
Seoud *et al.* (2015)	DR grading in 3 NPDR classes	Features extracted from red lesion probability maps (location, size, probability) + random forests	- Messidor	acc: 74.1%
Mahendran and Dhanasekaran (2015)	DR grading in 2 NPDR classes: moderate, severe	SVM/PNN classifier (features extracted from EX, based on GLCM)	- Messidor (only 370 images)	- grading: acc: 94.76%-97.89%; - detection: sens: 96.64%-98.68%, spec: 98.46%-100%
Nayak *et al.* (2008)	DR grading: normal, NPDR, PDR	NN trained to grade DR (features extracted from the segmentations of vessels, EX, etc.)	- 140 images (90 training)	- grading: acc: 93%; - detection: sens: 90%; spec: 100%
Pires *et al.* (2017)	Referable DR detection	Mid-level representations: BossaNova and Fisher Vector (extension of classical BoVW) + SVM classifier (non lesion-based approach)	- DR2; - Messidor	AUC: 0.964
Bhaskaranand *et al.* (2016)	Referable DR detection	Pixel feature descriptors + supervised learning to classify lesions (MAs, HEMs, EXs, CWS, IRMAs and NV) + supervised learning ensemble classifier	- 40532 images from the EyePACS	sens: 90%, spec: 63.2%
Seoud *et al.* (2016)	DR detection/referable	Dynamic shape features + random forest classifier to distinguish lesion from non lesion; image probability = max(prob(all lesions)) (considers only red lesions)	- Messidor; - Erlangen; - CARA1006 (private)	Messidor: - detection: AUC: 0.899, sens: 93.9%, spec: 50%; - referable: AUC: 0.916, sens: 96.2%, spec: 50%
Sánchez *et al.* (2011)	DR detection/referable	kNN classifier trained with features extracted from red and bright lesions (probabilities)	- Messidor	- detection: AUC: 0.876, sens: 92.2%, spec: 50%; - referable: AUC: 0.91, sens: 94.4%, spec: 50%
Casanova *et al.* (2014)	DR detection (and DR risk assessment)	Random forests (/logistic regression)	- 3443 participants from ACCORD-eye study; 4-year follow-up for 2856 patients	-
Antal and Hajdu (2012)	DR detection	Ensemble-based MA detection; DR detection based on MA presence	- Messidor	AUC: 0.90

References

Grading Diabetic Retinopathy from Stereoscopic Color Fundus Photographs—An Extension of the Modified Airlie House Classification. ETDRS Report Number 10. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology*, 98(5 Suppl):786–806, 1991. doi: 10.1016/S0161-6420(13)38012-9.

Abràmoff, M. D., Folk, J. C., Han, D. P., Walker, J. D., Williams, D. F., *et al.* Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmology*, 131(3): 351–357, 2013. doi: 10.1001/jamaophthalmol.2013.1743.

Abràmoff, M. D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., *et al.* Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Investigative Opthalmology & Visual Science*, 57(13):5200, 2016. doi: 10.1167/iovs.16-19964.

Abràmoff, M., Garvin, M., and Sonka, M. Retinal Imaging and Image Analysis. *IEEE Reviews in Biomedical Engineering*, 3:169–208, 2010. doi: 10.1109/RBME.2010.2084567.

Abràmoff, M. D. and Suttorp-Schulten, M. S. Web-based screening for diabetic retinopathy in a primary care population: The EyeCheck Project. *Telemedicine Journal and e-Health*, 11(6): 668–674, 2005. doi: 10.1089/tmj.2005.11.668.

Acharya, U. R., Ng, E. Y. K., Tan, J. H., Sree, S. V., and Ng, K. H. An integrated index for the identification of diabetic retinopathy stages using texture parameters. *Journal of Medical Systems*, 36(3):2011–2020, 2012. doi: 10.1007/s10916-011-9663-8.

ADCIS. Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (MESSIDOR). `http://www.adcis.net/en/Download-Third-Party/Messidor.html`.

Aiello, L. P., Cahill, M. T., and Wong, J. S. Systemic considerations in the management of diabetic retinopathy. *American Journal of Ophthalmology*, 132(5):760–776, 2001. doi: 10.1016/S0002-9394(01)01124-2.

Akram, M. U., Khalid, S., Tariq, A., Khan, S. A., and Azam, F. Detection and classification of retinal lesions for grading of diabetic retinopathy. *Computers in Biology and Medicine*, 45(1): 161–171, 2014. doi: 10.1016/j.compbiomed.2013.11.014.

Al-Bander, B., Al-Nuaimy, W., Williams, B. M., and Zheng, Y. Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. *Biomedical Signal Processing and Control*, 40:91–101, 2018. doi: 10.1016/j.bspc.2017.09.008.

Al Hajj, H., Lamard, M., Conze, P. H., Roychowdhury, S., Hu, X., *et al.* CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. *Medical Image Analysis*, 52:24–41, 2019. doi: 10.1016/j.media.2018.11.008.

American Academy of Ophthalmology. International clinical diabetic retinopathy disease severity scale. *American Academy of Ophthalmology*, (October), 2002.

Amin, J., Sharif, M., Yasmin, M., Ali, H., and Fernandes, S. L. A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions. *Journal of Computational Science*, 19:153–164, 2017. doi: 10.1016/j.jocs.2017.01.002.

Antal, B. and Hajdu, A. An Ensemble-Based System forMicroaneurysm Detection and Diabetic Retinopathy Grading. *IEEE Transactions on Biomedical Engineering*, 5(1):70–77, 2012. doi: 10.1109/TBME.2012.2193126.

Aquino, A. Establishing the macular grading grid by means of fovea centre detection using anatomical-based and visual-based features. *Computers in Biology and Medicine*, 55:61–73, 2014. doi: 10.1016/j.compbiomed.2014.10.007.

Aquino, A., Gegúndez-arias, M. E., and Marín, D. Detecting the Optic Disc Boundary in Digital Fundus Feature Extraction Techniques. *IEEE Transactions on Medical Imaging*, 29(11):1860–1869, 2010. doi: 10.1109/TMI.2010.2053042.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., *et al.* Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS ONE*, 12(6):e0177544, 2017a. doi: 10.1371/journal.pone.0177544.

Araújo, T., Aresta, G., Galdran, A., Costa, P., Mendonça, A. M., *et al.* UOLO - Automatic object detection and segmentation in biomedical images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11045 LNCS:165–173, 2018a. doi: 10.1007/978-3-030-00889-5_19.

Araújo, T., Wanderley, D., Carvalho, C., Costa, P., Maia, C., *et al.* Performance analysis of specialists and automatic methods in diabetic retinopathy screening. In *19th EURETINA, European Society of Retina Specialists*, Paris, 2019a.

Araújo, T., Maia, C., Penas, S., Mendonça, L., Carneiro, , *et al.* Explainable and uncertainty-aware system for diabetic retinopathy grading. In *20th EURETINA, European Society of Retina Specialists*, 2020a.

Araújo, T., Aresta, G., Almada-Lobo, B., Mendonça, A. M., and Campilho, A. Improving Convolutional Neural Network Design via Variable Neighborhood Search. In *ICIAR 2017 - International Conference on Image Analysis and Recognition LNCS 10317*, pages 371–379. 2017b. doi: 10.1007/978-3-319-59876-5_41.

Araújo, T., Mendonça, A. M., and Campilho, A. Parametric model fitting-based approach for retinal blood vessel caliber estimation in eye fundus images. *PLoS ONE*, 13(4):1–27, 2018b. doi: 10.1371/journal.pone.0194702.

Araújo, T., Aresta, G., Galdran, A., Costa, P., Mendonça, A. M., *et al.* UOLO: a framework for object detection and segmentation in medical images. In *Portuguese Conference on Pattern Recognition (RECPAD)*, Porto, 2019b.

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., *et al.* Data Augmentation for Improving Proliferative Diabetic Retinopathy Detection in Eye Fundus Images. *IEEE Access*, 8:182462–182474, 2020b. doi: 10.1109/ACCESS.2020.3028960.

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., *et al.* DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715, 7 2020c. doi: 10.1016/j.media.2020.101715.

Aresta, G., **Araújo, T.**, Campilho, A., Eloy, C., Polónia, A., *et al.* BACH: Grand challenge on breast cancer histology images. In *Proceedings of Machine Learning Research - MIDL 2019*, London, 2019a.

Aresta, G., **Araújo, T.**, Campilho, A., Eloy, C., Polónia, A., *et al.* BACH: Grand challenge on breast cancer histology images. In *Encontro Nacional de Jovens Investigadores em Oncologia*, Porto, 2019b.

Aresta, G., **Araújo, T.**, Kwok, S., Chennamsetty, S. S., Safwan, M., *et al.* BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019c. doi: 10.1016/j.media.2019.05.010.

Arévalo, J. F., Lasave, A. F., Zeballos, D. G., and Bonafonte-Royo, S. Diabetic Retinopathy. In *Retinal and Choroidal Manifestations of Selected Systemic Diseases*, pages 387–416. Springer New York, New York, NY, 2013. doi: 10.1007/978-1-4614-3646-1_21.

Arya, M., Sorour, O., Chaudhri, J., Alibhai, Y., Waheed, N. K., *et al.* Distinguishing Intraretinal Microvascular Abnormalities From Retinal Neovascularization Using Optical Coherence Tomography Angiography. *Retina*, 40(9):1686–1695, 2020. doi: 10.1097/IAE.0000000000002671.

Ayhan, M. S. and Berens, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning (MIDL)*, pages 1–9, 2018.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140.

Bellemo, V., Burlina, P., Yong, L., Wong, T. Y., and Ting, D. S. W. Generative Adversarial Networks (GANs) for Retinal Fundus Image Synthesis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11367 LNCS, pages 289–302, 2019. doi: 10.1007/978-3-030-21074-8_24.

Bhaskaranand, M., Ramachandra, C., Bhat, S., Cuadros, J., Nittala, M. G., *et al.* Automated Diabetic Retinopathy Screening and Monitoring Using Retinal Fundus Image Analysis. *Journal of Diabetes Science and Technology*, 10(2):254–261, 2016. doi: 10.1177/1932296816628546.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. doi: 10.1016/j.neunet.2018.07.011.

Casanova, R., Saldana, S., Chew, E. Y., Danis, R. P., Greven, C. M., *et al.* Application of random forests methods to diabetic retinopathy classification analyses. *PLoS ONE*, 9(6):1–8, 2014. doi: 10.1371/journal.pone.0098587.

Cheng, J., Liu, J., Tao, D., Yin, F., Wong, D. W. K., *et al.* Superpixel Classification Based Optic Cup Segmentation. *IEEE Transactions on Medical Imaging*, 32(6):1019–1032, 2013. doi: 10.1007/978-3-642-40760-4_53.

Combalia, M., Hueto, F., Puig, S., Malvehy, J., and Vilaplana, V. Uncertainty estimation in deep neural networks for dermoscopic image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 3211–3220, Seatle, 2020. doi: 10.1109/CVPRW50498.2020.00380.

Costa, P. and Campilho, A. Convolutional bag of words for diabetic retinopathy detection from eye fundus images. *IPSJ Transactions on Computer Vision and Applications*, 9(1):10, 2017. doi: 10.1186/s41074-017-0023-6.

Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abràmoff, M., *et al.* End-to-End Adversarial Retinal Image Synthesis. *IEEE Transactions on Medical Imaging*, 37(3):781–791, 2018. doi: 10.1109/TMI.2017.2759102.

Costa, P., **Araújo, T.**, Aresta, G., Galdran, A., Mendonça, A. M., *et al.* EyeWeS: Weakly Supervised Pre-Trained Convolutional Neural Networks for Diabetic Retinopathy Detection. In *International Conference on Machine Vision Applications (MVA)*, pages 5–15, Tokyo, 2019. doi: 10.23919/MVA.2019.8757991.

Cuadros, J. and Bresnick, G. EyePACS: An adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3):509–516, 2009. doi: 10.1177/193229680900300315.

Dai, B., Wu, X., and Bu, W. Optic disc segmentation based on variational model with multiple energies. *Pattern Recognition*, 64:226–235, 2017. doi: 10.1016/j.patcog.2016.11.017.

Dai, P., Sheng, H., Zhang, J., Li, L., Wu, J., *et al.* Retinal Fundus Image Enhancement Using the Normalized Convolution and Noise Removing. *International Journal of Biomedical Imaging*, 2016:1–12, 2016. doi: 10.1155/2016/5075612.

Dashtbozorg, B., Mendonça, A., and Campilho, A. Optic disc segmentation using the sliding band filter. *Computers in Biology and Medicine*, 56:1–12, 2015. doi: 10.1016/j.compbiomed.2014.10.009.

de la Torre, J., Valls, A., and Puig, D. A Deep Learning Interpretable Classifier for Diabetic Retinopathy Disease Grading. *arXiv pre-print*, 1712.08107, 2017.

de la Torre, J., Puig, D., and Valls, A. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154, 2018. doi: 10.1016/j.patrec.2017.05.018.

Decenciere, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., *et al.* Feedback on a publicly distributed image database: The Messidor database. *Image Analysis and Stereology*, 33(3):231–234, 2014. doi: 10.5566/ias.1155.

Direção Geral da Saúde. Ratreio da Retinopatia Diabética. *Número 016/2018*, pages 1–13, 2018.

Dutra Medeiros, M., Mesquita, E., Gardete-Correia, L., Moita, J., Genro, V., *et al.* First Incidence and Progression Study for Diabetic Retinopathy in Portugal, the RETINODIAB Study: Evaluation of the Screening Program for Lisbon Region. *Ophthalmology*, 122(12):2473–2481, 2015. doi: 10.1016/j.ophtha.2015.08.004.

Fadzil, M. H. A., Izhar, L. I., Nugroho, H., and Nugroho, H. A. Analysis of retinal fundus images for grading of diabetic retinopathy severity. *Medical and Biological Engineering and Computing*, 49(6):693–700, 2011. doi: 10.1007/s11517-011-0734-2.

Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G., Kenton, Z., *et al.* A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. In *4th workshop on Bayesian Deep Learning (NeurIPS)*, 2019.

Friston, K. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010. doi: 10.1038/nrn2787.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, New York, 2016. doi: 10.1109/TKDE.2015.2507132.

Galdran, A., **Araújo, T.**, Mendonça, A. M., and Campilho, A. Retinal Image Quality Assessment by Mean-Subtracted Contrast-Normalized Coefficients. In *VipIMAGE 2017*, pages 844–853, Cham, 2018a. Springer International Publishing. doi: 10.1007/978-3-319-68195-5_92.

Galdran, A., Costa, P., Bria, A., **Araújo, T.**, Mendonça, A. M., *et al.* A No-Reference Quality Metric for Retinal Vessel Tree Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 82–90, Cham, 2018b. Springer International Publishing. doi: 978-3-030-00928-1_10.

Galdran, A., Chelbi, J., Kobi, R., Dolz, J., Lombaert, H., *et al.* Non-uniform label smoothing for diabetic retinopathy grading from retinal fundus images with deep neural networks. *Translational Vision Science and Technology*, 9(2):1–8, 2020a. doi: 10.1167/tvst.9.2.34.

Galdran, A., Dolz, J., Chakor, H., Lombaert, H., and Ben Ayed, I. Cost-Sensitive Regularization for Diabetic Retinopathy Grading from Eye Fundus Images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12265 LNCS:665–674, 2020b. doi: 10.1007/978-3-030-59722-1_64.

Garg, S. and Davis, R. M. Diabetic retinopathy screening update. *Clinical Diabetes*, 27(4):140–145, 2009. doi: 10.2337/diaclin.27.4.140.

Gargeya, R. and Leng, T. Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology*, 124(7):962–969, 7 2017. doi: 10.1016/j.ophtha.2017.02.008.

Gegundez-Arias, M. E., Marin, D., Bravo, J. M., and Suero, A. Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques. *Computerized Medical Imaging and Graphics*, 37(5-6):386–393, 2013. doi: 10.1016/j.compmedimag.2013.06.002.

Ghasemi Falavarjani, K., Tsui, I., and Sadda, S. R. Ultra-wide-field imaging in diabetic retinopathy. *Vision Research*, 139:187–190, 2017. doi: 10.1016/j.visres.2017.02.009.

González-Gonzalo, C., Liefers, B., Ginneken, B. V., and Sánchez, C. I. Improving weakly-supervised lesion localization with iterative saliency map refinement. 2018.

Gonzalez-Gonzalo, C., Liefers, B., van Ginneken, B., and Sanchez, C. I. Iterative Augmentation of Visual Evidence for Weakly-Supervised Lesion Localization in Deep Interpretability Frameworks: Application to Color Fundus Images. *IEEE transactions on medical imaging*, 39(11): 3499–3511, 2020. doi: 10.1109/TMI.2020.2994463.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR)*, pages 1–11, 2015.

Graves, A. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2348–2356, 2011.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama*, 304(6):649–656, 2016. doi: 10.1001/jama.2016.17216.

Guo, Y., Wang, K., Yang, S., Wang, Y., Gao, P., *et al.* Structure-Aware Noise Reduction Generative Adversarial Network for Optical Coherence Tomography Image. In *Ophthalmic Medical Image Analysis (OMIA)*, pages 9–17, Cham, 2019. Springer International Publishing. doi: 10.1007/978-3-030-32956-3_2.

Gupta, G., Kulasekaran, S., Ram, K., Joshi, N., Sivaprakasam, M., *et al.* Local characterization of neovascularization and identification of proliferative diabetic retinopathy in retinal fundus images. *Computerized Medical Imaging and Graphics*, 55:124–132, 2017. doi: 10.1016/j.compmedimag.2016.08.005.

Hall, A. B. Recognising and managing diabetic retinopathy How I look for diabetic retinopathy: a vision technician ' s experience. *Community Eye Health*, 24(75), 2011.

Hand, D. J. and Till, R. J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2):171–186, 2001. doi: 10.1023/A:1010920819831.

Harangi, B. and Hajdu, A. Improving automatic exudate detection based on the fusion of the results of multiple active contours. *Proceedings - International Symposium on Biomedical Imaging*, pages 45–48, 2013. doi: 10.1109/ISBI.2013.6556408.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2980-2988, Venice, 2017. doi: 10.1109/ICCV.2017.322.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, 2016. doi: 10.1109/CVPR.2016.90.

Hernández-Lobato, J. M. and Adams, R. P. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *32nd International Conference on Machine Learning (ICML)*, volume 37, pages 1861–1869, Lille, 2015.

Hipwell, J. Ophthalmic Photography: Retinal Photography, Angiography, and Electronic Imaging. *British Journal of Ophthalmology*, 81, 1997.

Huang, G., Liu, Z., Maaten, L. v. d., and Weinberger, K. Q. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269. IEEE, 2017. doi: 10.1109/CVPR.2017.243.

Ioffe, S. and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 2015. PMLR.

Kamble, R., Kokare, M., Deshmukh, G., Hussin, F. A., and Mériaudeau, F. Localization of optic disc and fovea in retinal images using intensity based line scanning analysis. *Computers in Biology and Medicine*, 87:382–396, 2017. doi: 10.1016/j.compbiomed.2017.04.016.

Kar, S. S. and Maity, S. P. Automatic Detection of Retinal Lesions for Screening of Diabetic Retinopathy. *IEEE Transactions on Biomedical Engineering*, 65(3):608–618, 2018. doi: 10.1109/TBME.2017.2707578.

Kendall, A. and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NIPS)*, pages 5575–5585, Long Beach, 2017.

Kendall, M. G. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 6 1938. doi: 10.1093/biomet/30.1-2.81.

Köhler, T., Budai, A., Kraus, M. F., Odstrčilik, J., Michelson, G., *et al.* Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *Proceedings of CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 95–100, 2013. doi: 10.1109/CBMS.2013.6627771.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., *et al.* Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018. doi: 10.1016/j.ophtha.2018.01.034.

Krizhevsky, A., Sutskever, I., and Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NIPS)*, 25:1106–1114, 2012.

Kruskal, W. H., Wallis, W. A., Kruskal, W. H., and Wallis, W. A. Journal of the American Statistical Analysis. *Journal of the American Statistical Association*, 47(260):583–621, 2012. doi: 10.2307/2280779.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6405–6416, 2017.

Laves, M.-H., Fast, J. F., Kahrs, L. A., and Ortmaier, T. Well-Calibrated Regression Uncertainty in Medical Imaging with Deep Learning. *Proceedings of Machine Learning Research*, 121(20): 393–412, 2020.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):1–14, 2017. doi: 10.1038/s41598-017-17876-z.

Li, H. and Chutatape, O. Automatic location of optic disk in retinal images. In *IEEE International Conference on Image Processing*, volume 2, pages 837–840, 2001. doi: 10.1109/icip.2001.958624.

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2017. doi: 10.1109/TPAMI.2018.2858826.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., *et al.* A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017. doi: 10.1016/j.media.2017.07.005.

Liu, T. Y. A. and Arevalo, J. F. Wide-field imaging in proliferative diabetic retinopathy. *International Journal of Retina and Vitreous*, 5(Suppl 1):20, 12 2019. doi: 10.1186/s40942-019-0170-2.

MacKay, D. J. *Bayesian methods for adaptive models (Ph.D. Dissertation)*. PhD thesis, California Institute of Technology, 1992.

Mahendran, G. and Dhanasekaran, R. Investigation of the severity level of diabetic retinopathy using supervised classifier algorithms. *Computers and Electrical Engineering*, 45:312–323, 2015. doi: 10.1016/j.compeleceng.2015.01.013.

Massin, P., Chabouis, A., Erginay, A., Viens-Bitker, C., Lecleire-Collet, A., *et al.* OPHDIAT©: A telemedical network screening system for diabetic retinopathy in the Île-de-France. *Diabetes and Metabolism*, 34(3):227–234, 2008. doi: 10.1016/j.diabet.2007.12.006.

Mendonça, A. M., Melo, T., **Araújo, T.**, and Campilho, A. Optic Disc and Fovea Detection in Color Eye Fundus Images. In *Image Analysis and Recognition*, pages 332–343, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-50516-5_29.

Menti, E., Bonaldi, L., Ballerini, L., Ruggeri, A., and Trucco, E. Simulation and Synthesis in Medical Imaging. In *Simulation and Synthesis in Medical Imaging - First International Workshop, SASHIMI 2016 Held in Conjunction with MICCAI 2016 Proceedings*, volume 9968 of *Lecture Notes in Computer Science*, pages 167–176, Athens, 2016. Springer International Publishing. doi: 10.1007/978-3-319-46630-9.

Mookiah, M. R. K., Acharya, U. R., Martis, R. J., Chua, C. K., Lim, C. M., *et al.* Evolutionary algorithm based classifier parameter tuning for automatic diabetic retinopathy grading: A hybrid feature extraction approach. *Knowledge-Based Systems*, 39:9–22, 2013. doi: 10.1016/j.knosys.2012.09.008.

Morales, S., Naranjo, V., Angulo, U., and Alcaniz, M. Automatic detection of optic disc based on PCA and mathematical morphology. *IEEE Transactions on Medical Imaging*, 32(4):786–796, 2013. doi: 10.1109/TMI.2013.2238244.

National Center for Biotechnology Information. PubMed. `https://pubmed.ncbi.nlm.nih.gov/`, 2020.

Nayak, J., Bhat, P. S., Acharya U, R., Lim, C. M., and Kagathi, M. Automated identification of diabetic retinopathy stages using digital fundus images. *Journal of Medical Systems*, 32(2):107–115, 2008. doi: 10.1007/s10916-007-9113-9.

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, 1996. doi: 10.1007/978-1-4612-0745-0.

Nguyen, T. T. and Wong, T. Y. Retinal vascular changes and diabetic retinopathy. *Current Diabetes Reports*, 9(4):277–283, 2009. doi: 10.1007/s11892-009-0043-4.

Niemeijer, M., Van Ginneken, B., Staal, J., Suttorp-Schulten, M. S. A., and Abràmoff, M. D. Automatic detection of red lesions in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 24(5):584–592, 2005. doi: 10.1109/TMI.2005.843738.

Niemeijer, M., Abràmoff, M. D., and van Ginneken, B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. *Medical Image Analysis*, 10(6):888–898, 2006. doi: 10.1016/j.media.2006.09.006.

Niemeijer, M., Ginneken, B. v., Russell, S. R., Suttorp-Schulten, M. S., and Abramoff, M. D. Automated Detection and Differentiation of Drusen, Exudates, and Cottonwool Spots in Digital Color Fundus Photographs for Early Diagnosis of Diabetic Retinopathy. *Invest Ophthalmol Vis Sci*, 48(5):2260–7, 2007. doi: 10.1167/iovs.06-0996.

Pedrosa, M., Silva, J. M., Silva, J. F., Matos, S., and Costa, C. SCREEN-DR: Collaborative platform for diabetic retinopathy. *International Journal of Medical Informatics*, 120:137–146, 12 2018. doi: 10.1016/j.ijmedinf.2018.10.005.

Peterson, C. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1: 995–1019, 1987.

Pires, R., Avila, S., Jelinek, H. F., Wainer, J., Valle, E., *et al.* Beyond Lesion-Based Diabetic Retinopathy: A Direct Approach for Referral. *IEEE Journal of Biomedical and Health Informatics*, 21(1):193–200, 2017. doi: 10.1109/JBHI.2015.2498104.

Pires, R., Jelinek, H. F., Wainer, J., and Rocha, A. Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection. In *25th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 229–236, Ouro Preto, 8 2012. IEEE. doi: 10.1109/SIBGRAPI.2012.39.

Pires Dias, J. M., Oliveira, C. M., and Da Silva Cruz, L. A. Retinal image quality assessment using generic image quality indicators. *Information Fusion*, 19(1):73–90, 2014. doi: 10.1016/j.inffus.2012.08.001.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., *et al.* Indian Diabetic Retinopathy Image Dataset (IDRiD). *IEEE Dataport*, 2018. doi: 10.21227/H25W98.

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., *et al.* IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge. *Medical Image Analysis*, 59:101561, 1 2020. doi: 10.1016/j.media.2019.101561.

Prentasic, P. and Loncaric, S. Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion. *computer methods and programs in biomedicine*, 137:281–292, 2016. doi: 10.1109/ISPA.2015.7306056.

Pujitha, A. K. and Sivaswamy, J. Retinal Image Synthesis for CAD development. In *15th International Conference on Image Analysis and Recognition, ICIAR*, volume 10882, page 613–6211, 2018. doi: 10.1007/978-3-319-93000-8.

Quelhas, P., Marcuzzo, M., Mendonça, A. M., and Campilho, A. Cell nuclei and cytoplasm joint segmentation using the sliding band filter. *IEEE Transactions on Medical Imaging*, 29(8): 1463–1473, 2010. doi: 10.1109/TMI.2010.2048253.

Quellec, G., Charrière, K., Boudi, Y., Cochener, B., and Lamard, M. Deep image mining for diabetic retinopathy screening. *Medical Image Analysis*, 39:178–193, 2017. doi: 10.1016/j.media.2017.04.012.

Quellec, G., Russell, S. R., and Abramoff, M. D. Optimal filter framework for automated, instantaneous detection of lesions in retinal images. *IEEE Transactions on Medical Imaging*, 30(2): 523–533, 2011. doi: 10.1109/TMI.2010.2089383.

Redmon, J. and Farhadi, A. YOLO9000: Better, Faster, Stronger. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, Honolulu, 2017. doi: 10.1109/CVPR.2017.690.

Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

Reza, A. W., Eswaran, C., and Dimyati, K. Diagnosis of diabetic retinopathy: Automatic extraction of optic disc and exudates from retinal images using marker-controlled watershed transformation. *Journal of Medical Systems*, 35(6):1491–1501, 2011. doi: 10.1007/s10916-009-9426-y.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, 9351:234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Roy, P., Tennakoon, R., Cao, K., Sedai, S., Mahapatra, D., *et al.* A novel hybrid approach for severity assessment of Diabetic Retinopathy in colour fundus images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 1078–1082. IEEE, 4 2017. doi: 10.1109/ISBI.2017.7950703.

Roychowdhury, S., Koozekanani, D., and Parhi, K. DREAM: Diabetic Retinopathy Analysis using Machine Learning. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1717–1728, 2014. doi: 10.1109/JBHI.2013.2294635.

Roychowdhury, S., Koozekanani, D. D., Kuchinka, S. N., and Parhi, K. K. Optic disc boundary and vessel origin segmentation of fundus images. *IEEE Journal of Biomedical and Health Informatics*, 20(6):1562–1574, 2016. doi: 10.1109/JBHI.2015.2473159.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Salazar-Gonzalez, A., Kaba, D., Li, Y., and Liu, X. Segmentation of Blood Vessels and Optic Disc in Retinal Images. *IEEE Journal of Biomedical and Health Informatics*, 18(6):1874–1886, 2014. doi: 10.1109/JBHI.2014.2302749.

Saleh, M. D. and Eswaran, C. An automated decision-support system for non-proliferative diabetic retinopathy disease based on MAs and HAs detection. *Computer Methods and Programs in Biomedicine*, 108(1):186–196, 2012. doi: 10.1016/j.cmpb.2012.03.004.

Samek, W., Binder, A., Montavon, G., Bach, S., and Müller, K.-R. Evaluating the visualization of what a Deep Neural Network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660 – 2673, 2016. doi: 10.1109/TNNLS.2016.2599820.

Sánchez, C. I., Niemeijer, M., Dumitrescu, A. V., Suttorp-Schulten, M. S. A., Abràmoff, M. D., *et al.* Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. *Investigative ophthalmology & visual science*, 52:4866–4871, 2011. doi: 10.1167/iovs.10-6633.

Sawilowsky, S. S. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599, 2009. doi: 10.22237/jmasm/1257035100.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., *et al.* Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*, 126(4):552–564, 2019. doi: 10.1016/j.ophtha.2018.11.016.

Schully, T. Diabetes in Numbers. *Nature*, 485(7398):S2–S3, 2012. doi: 10.1038/nn.2369.

Scotland, G. S., McNamee, P., Philip, S., Fleming, A. D., Goatman, K. A., *et al.* Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in Scotland. *The British journal of ophthalmology*, 91(11):1518–23, 2007. doi: 10.1136/bjo.2007.120972.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, 2017. IEEE. doi: 10.1109/ICCV.2017.74.

Seoud, L., Hurtut, T., Chelbi, J., Cheriet, F., and Langlois, J. P. Red Lesion Detection using Dynamic Shape Features for Diabetic Retinopathy Screening. *IEEE Transactions on Medical Imaging*, 35(4):1116–1126, 2016. doi: 10.1109/TMI.2015.2509785.

Seoud, L., Chelbi, J., and Cheriet, F. Automatic grading of diabetic retinopathy on a public database. In *Proceedings of the Ophthalmic Medical Image Analysis Second International Workshop, OMIA 2015, Held in Conjunction with MICCAI 2015*, pages 97–104, Munich, 2015.

Sevik, U., Köse, C., Berber, T., and Erdöl, H. Identification of suitable fundus images using automated quality assessment methods. *Journal of Biomedical Optics*, 19(4):046006, 2014. doi: 10.1117/1.jbo.19.4.046006.

Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICRL)*, pages 1–14, 2015. doi: 10.1016/j.infsof.2008.09.005.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICRL)*, Calgary, 2014.

Sociedade Portuguesa de Diabetologia. *Diabetes Factos e Números - O ano de 2015: Relatório Anual do Observatório Nacional da Diabetes*. 2016.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for Simplicity: The All Convolutional Net. In *3rd International Conference on Learning Representations (ICLR)*, pages 1–14, San Diego, 2015.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

Staal, J. J., Abràmoff, M. D., Niemeijer, M., Viergever, M. a., and Van Ginneken, B. Ridge based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23 (4):501–509, 2004. doi: 10.1109/TMI.2004.825627.

Sullivan, G. M. and Feinn, R. Using Effect Size—or Why the P Value Is Not Enough . *Journal of Graduate Medical Education*, 4(3):279–282, 2012. doi: 10.4300/jgme-d-12-00156.1.

Sun, J. K., Silva, P. S., Cavallerano, J. D., Blodi, B. A., Davis, M. D., *et al.* Retinal Vascular Disease. In *Ryan's Retina - Volume 2*, chapter 51 - Proli, pages 1091–1128. Elsevier, 6th edition, 2018.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning (ICML)*, volume 70, page 3319–3328, 2017.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., *et al.* Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, 2015. IEEE. doi: 10.1109/CVPR.2015.7298594.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. doi: 10.1002/2014GB005021.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 4278–4284, 2017.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., *et al.* Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016. doi: 10.1109/TMI.2016.2535302.

Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., and Kawashima, H. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. *Plos One*, 12 (6):e0179790, 2017. doi: 10.1371/journal.pone.0179790.

The Diabetic Retinopathy Study Research Group. Four Risk Factors for Severe Visual Loss in Diabetic Retinopathy: The Third Report From the Diabetic Retinopathy Study. *Archives of Ophthalmology*, 97(4):654–655, 1979. doi: 10.1001/archopht.1979.01020010310003.

The Royal College of Ophthalmologists. Diabetic Retinopathy Guidelines. Technical Report December, 2012.

Voets, M., Møllersen, K., and Bongo, L. A. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS ONE*, 14(6):1–11, 2019. doi: 10.1371/journal.pone.0217541.

Wanderley, D. S., **Araújo, T.**, Carvalho, C. B., Maia, C., Penas, S., *et al.* Analysis of the performance of specialists and an automatic algorithm in retinal image quality assessment. In *IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, Lisboa, 2019. IEEE. doi: 10.1109/ENBENG.2019.8692506.

Wickstrøm, K., Kampffmeyer, M., and Jenssen, R. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60:101619, 2020. doi: 10.1016/j.media.2019.101619.

Wild. Estimates for the year 2000 and projections for 2030. *World Health*, 27(5):1047–1053, 2004. doi: 10.2337/diacare.27.5.1047DiabetesCareMay2004vol.27no.51047-1053.

Wilkinson, C. P., Ferris, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., *et al.* Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003. doi: 10.1016/S0161-6420(03)00475-5.

Wu, B., Zhu, W., Shi, F., Zhu, S., and Chen, X. Automatic detection of microaneurysms in retinal fundus images. *Computerized Medical Imaging and Graphics*, 55:106–112, 2017. doi: 10.1016/j.compmedimag.2016.08.001.

Wu, L., Fernandez-Loaiza, P., Sauma, J., Hernandez-Bogantes, E., and Masis, M. Classification of diabetic retinopathy and diabetic macular edema. *World journal of diabetes*, 4(6):290–4, 2013. doi: 10.4239/wjd.v4.i6.290.

Xie, S., Girshick, R., and Doll, P. Aggregated Residual Transformations for Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, Honolulu, 2017. doi: 10.1109/CVPR.2017.634.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pages 3320–3328, 2014.

Yu, H., Barriga, E. S., Agurto, C., Echegaray, S., Pattichis, M. S., *et al.* Fast Localization and Segmentation of Optic Disk in Retinal Images Using Directional Matched Filtering and Level Sets. *IEEE Transactions on Information Technology in Biomedicine*, 16(4):644–657, 7 2012. doi: 10.1109/TITB.2012.2198668.

Yun, W. L., Rajendra Acharya, U., Venkatesh, Y. V., Chee, C., Min, L. C., *et al.* Identification of different stages of diabetic retinopathy using retinal optical images. *Information Sciences*, 178 (1):106–121, 2008. doi: 10.1016/j.ins.2007.07.020.

Zeiler, M. D. and Fergus, R. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833, Cham, 2014. Springer International Publishing. doi: 10.1007/978-3-319-10590-1_53.

Zhang, B., Wu, X., You, J., Li, Q., and Karray, F. Detection of microaneurysms using multi-scale correlation coefficients. *Pattern Recognition*, 43(6):2237–2248, 2010.

Zhang, X., Thibault, G., Decencière, E., Marcotegui, B., Laÿ, B., *et al.* Exudate detection in color retinal images for mass screening of diabetic retinopathy. *Medical Image Analysis*, 18(7): 1026–1043, 2014. doi: 10.1016/j.media.2014.05.004.

Zhao, H., Li, H., Maurer-Stroh, S., and Cheng, L. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical Image Analysis*, 49:14–26, 2018. doi: 10.1016/j.media.2018.07.001.

Zhao, Z., Zhang, K., Hao, X., Tian, J., Chin, M., *et al.* BIRA-NET : Bilinear Attention Net for Diabetic Retinopathy Grading. In *IEEE International Conference on Image Processing (ICIP)*, pages 1385–1389, Taipei, 2019. IEEE. doi: 10.1109/ICIP.2019.8803074.