UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

GABRIEL TOZATTO ZAGO

Diabetic Retinopathy Detection Based On Deep Learning

Vitória
2019

GABRIEL TOZATTO ZAGO

Diabetic Retinopathy Detection Based On Deep Learning

Tese de doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica, do Centro Tecnológico da Universidade Federal do Espírito Santo, como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica.

Universidade Federal do Espírito Santo

Centro Tecnológico

Programa de Pós-graduação em Engenharia Elétrica

Orientador: Prof. Dr. Rodrigo Varejão Andreão

Coorientador: Prof. Dr. Evandro Ottoni Teatini Salles

Vitória

2019

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Zago, Gabriel Tozatto, 1988-
Z18d Diabetic retinopathy detection based on deep learning / Gabriel Tozatto Zago. - 2019.
96 f. : il.

Orientador: Rodrigo Varejão Andreão.
Coorientador: Evandro Ottoni Teatini Salles.
Tese (Doutorado em Engenharia Elétrica) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Retina - Doenças. 2. Aprendizado do computador. 3. Redes neurais (Computação). 4. Retina. 5. Processamento de imagens. I. Andreão, Rodrigo Varejão. II. Salles, Evandro Ottoni Teatini. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 621.3

GABRIEL TOZATTO ZAGO

Diabetic Retinopathy Detection Based on Deep Learning

Tese de doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como pré-requisito para obtenção do título de doutor.

Trabalho aprovado, Brasil, 20 de dezembro de 2019:

Rodrigo Varejão Andreão – Ifes
Orientador

Evandro Ottoni Teatini Salles – UFES
Co-orientador

Aura Conci – UFF
Participante Externo

Mariana Rampinelli Fernandes – Ifes
Participante Externo

Thomas Walter Rauber – UFES
Participante Interno

Patrick Marques Ciarelli – UFES
Participante Interno

Brasil

2019

Dedico este trabalho à minha esposa, com admiração e gratidão por seu apoio, carinho e presença ao longo do período de elaboração deste trabalho.

AGRADECIMENTOS

A Deus, à Virgem Santíssima e a São José, que me levaram pelas mãos até a conclusão desta tese.

À minha esposa Rayhanne e a meus filhos, Bento e Maria Isabel, por serem minha principal motivação e não terem me deixado desistir.

Aos professores Rodrigo Varejão e Bernadette Dorrizi, pelas incontáveis conversas e paciência com minhas limitações.

Ao professor Evandro, que me aceitou como orientado mesmo num tema transversal e me deu dicas preciosas.

Magnificat anima mea Dominum,et exsultavit spiritus meus in Deo salvatore meo, quia respexit humilitatem ancillae suae. Ecce enim ex hoc beatam me dicent omnes generationes, quia fecit mihi magna, qui potens est, et sanctum nomen eius, et misericordia eius in progenies et progenies timentibus eum.

(Magnificat)

RESUMO

A detecção precoce de retinopatia diabética (RD) é essencial, pois o tratamento oportuno pode reduzir ou até impedir a perda da visão. Além disso, a localização automática das regiões da imagem da retina que podem conter lesões pode auxiliar os especialistas na tarefa de detecção da doença. Ao mesmo tempo, imagens de baixa qualidade não permitem um diagnóstico médico preciso e causam o inconveniente de o paciente ter de retornar ao centro médico para repetir o exame de fundo do olho.

Nesta tese, argumentamos que é possível propor um sistema com base na avaliação da qualidade da imagem e na localização de lesões vermelhas para detectar automaticamente RD com desempenho semelhante ao de especialistas, considerando que uma segmentação aproximada é suficiente para produzir um marcador discriminante de uma lesão.

Um sistema automático robusto é proposto para avaliar a qualidade das imagens de retina visando auxiliar os profissionais de saúde durante um exame de fundo de olho. Propomos uma rede neural convolucional (CNN) pré-treinada em imagens não médicas para extrair características gerais de imagem. Os pesos da CNN são ajustados através de um procedimento de ajuste fino, resultando em um classificador de bom desempenho ajustado com uma pequena quantidade de imagens rotuladas.

Também projetamos um modelo de localização de lesões usando uma abordagem de aprendizado profundo baseada em regiões. Nosso objetivo é reduzir a complexidade do modelo e melhorar seu desempenho. Para esse fim, desenvolvemos um procedimento (incluindo dois modelos de redes neurais convolucionais) para selecionar as regiões utilizadas no treinamento, de modo que os exemplos desafiadores recebessem atenção especial durante o processo de treinamento. Usando anotações de região, uma predição de RD pode ser definida na imagem inicial, sem a necessidade de treinamento especial. Nossa abordagem baseada em região permite que o modelo seja treinado com apenas 28 imagens, resultando em desempenho semelhante a trabalhos que usaram mais de um milhão de imagens rotuladas.

O desempenho da CNN para avaliação da qualidade foi medido através de dois bancos de dados publicamente disponíveis (DRIMDB e ELSA-Brasil) usando dois procedimentos diferentes: validação cruzada intra-banco de dados e inter-banco de dados. A CNN alcançou uma área sob a curva característica de operação do receptor (AUC) de 99,98% no DRIMDB e uma AUC de 98,56% no ELSA-Brasil no experimento interbancos (ou seja, com treinamento e testes não realizados no mesmo conjunto de dados). Esses resultados mostram a robustez do

modelo proposto para vários dispositivos de aquisição de imagens sem a necessidade de adaptação especial, tornando-o um bom candidato para uso em cenários operacionais.

O modelo de localização da lesão foi treinado no banco de dados de Retinopatia Diabética Padrão, Nível de Calibração 1 (DIARETDB1) e testado em vários bancos de dados (incluindo Messidor) sem qualquer adaptação adicional. Alcançou uma AUC de 0,912 - 95% IC 0,897-0,928 para a triagem de RD e uma sensibilidade de 0,940-95% CI 0,921-0,959. Esses valores são similares a outras abordagens do estado da arte.

Os resultados sugerem que a hipótese da proposta é confirmada.

Palavras-chave: Imagens de retina. Aprendizado profundo. Retinopatia diabética. Redes neurais convolucionais. Qualidade de imagem. Localização de lesão.

ABSTRACT

Detecting the early signs of diabetic retinopathy (DR) is essential, as timely treatment might reduce or even prevent vision loss. Moreover, automatically localizing the regions of the retinal image that might contain lesions can favorably assist specialists in the task of detection. At the same time, poor-quality retinal images do not allow an accurate medical diagnosis, and it is inconvenient for a patient to return to a medical center to repeat the fundus photography exam.

In this thesis, we argue that it is possible to propose a pipeline based on quality assessment and red lesion localization to achieve automatic DR detection with performance similar to experts considering that a rough segmentation is sufficient to produce a discriminant marker of a lesion.

A robust automatic system is proposed to assess the quality of retinal images aiming at assisting health care professionals during a fundus photograph exam. We propose a convolutional neural network (CNN) pretrained on non-medical images for extracting general image features. The weights of the CNN are further adjusted via a fine-tuning procedure, resulting in a performant classifier using only with a small quantity of labeled images.

We also designed a lesion localization model using a deep network patch-based approach. Our goal was to reduce the complexity of the implementation while improving its performance. For this purpose, we designed an efficient procedure (including two convolutional neural network models) for selecting the training patches, such that the challenging examples would be given special attention during the training process. Using the labeling of the region, a DR decision can be given to the initial image, without the need for special training. Our patch-based approach allows the model to be trained with only 28 images achieving similar results to works that used over a million of labeled images.

The CNN performance for quality assessment was evaluated on two publicly available databases (i.e., DRIMDB and ELSA-Brasil) using two different procedures: intra-database and inter-database cross-validation. The CNN achieves an area under the receiver operating characteristic curve (AUC) of 99.98% on DRIMDB and an AUC of 98.56% on ELSA-Brasil in the inter-database experiment, where training and testing were not performed on the same database. These results suggest the robustness of the proposed model to various image acquisitions without requiring special adaptation, thus making it a good candidate for use in operational clinical scenarios.

The lesion localization model was trained on the Standard Diabetic Retinopathy Database, Calibration Level 1 (DIARETDB1) database and was tested on several databases (including Messidor) without any further adaptation. It reaches an area under the receiver operating characteristic curve of 0.912 - 95%CI 0.897-0.928 for DR screening, and a sensitivity of 0.940-95%CI 0.921-0.959. These values are competitive with other state-of-the-art approaches.

The results suggest that the given hypothesis is confirmed.

Keywords: Retinal images. Deep learning. Diabetic retinopathy. Convolutional neural networks. Image quality. Lesion localization.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	artificial neural network
ASG	Automatic seed generation
AUC	Area Under the Curve
CADS	Computer-Assisted Diagnostic System
CI	confidence interval
CNN	Fully Convolutional Neural Networks
DIARETDB0	Standard Diabetic Retinopathy Database, Calibration Level 0
DIARETDB1	Standard Diabetic Retinopathy Database, Calibration Level 1
DM	Diabetes Mellitus
DM2	type II diabetes mellitus
DME	Diabetic Macular Edema
DR	Diabetic retinopathy
DRIMDB	Diabetic Retinopathy Image Database
ETDRS	Early Treatment Diabetic Retinopathy Study
EyePACS	Eye Picture Archive Communication System
FOV	Field Of View
FROC	Free-response receiver operating characteristic
GAN	Generative Adversarial Networks
HE	Hard exudates
HEM	Hemorrhages
IDRiD	Indian Diabetic Retinopathy Image Dataset
ISC	image structure clustering
k-NN	k-nearest neighbor
MA	Microaneurysms
ME	Macular Edema
MSCF	Multi-scale correlation filtering
NV	Neovascularization
PDR	Proliferative Diabetic Retinopathy
PDR	Proliferative Diabetic Retinopathy
PLS	Partial Least Square
ReLUs	Rectified linear units

RGB	Red, Green, and Blue
RIQA	Retinal Image Quality Assessment
RNFL	Retinal Nerve Fiber Layer
ROC	Retinopathy Online Challenge
ROI	Region Of Interest
SE	Soft exudates
Se	Sensitivity
Sp	Specificity
STARE	Structured Analysis of the Retina
STFM	Spatiotemporal feature map classifier
SVM	Support Vector Machine
WCSDR	Welsh Community Diabetic Retinopathy Study
WHO	World Health Organization

CONTENTS

1 INTRODUCTION

Diabetes mellitus (DM) is a chronic disease characterized by hyperglycemia (FAUST et al., 2012b). The World Health Organization (WHO) recognizes three significant forms of diabetes: type I, type II, and gestational diabetes, which have similar signs, symptoms, and consequences, but different causes and different distributions in the population (ALGHADYAN, 2011).

Currently, DM represents a severe public health problem due to the high prevalence in the world, especially in developing countries, due to morbidity and because it is one of the primary cardiovascular and cerebrovascular risk factors.

The appearance of chronic complications marks the natural evolution of diabetes, usually classified as microvascular: retinopathy (FAUST et al., 2012a; HARMAN-BOEHM et al., 2007), neuropathy (EWING; CLARKE, 1982), nephropathy (GUTHRIE; GUTHRIE, 2003), cardiomyopathy (GRUNDY et al., 1999), and macrovascular: coronary artery disease, cerebrovascular and peripheral vascular disease. All complications are responsible for: significant morbidity and mortality, with cardiovascular and renal mortality rates, blindness, limb amputation, and loss of function and quality of life lower than individuals without diabetes.

With the aging of the population and lifestyle increasingly characterized by physical inactivity and eating habits that predispose to the accumulation of body fat, the tendency is to increase the percentage of patients with type II diabetes mellitus (DM2) (HARNEY, 2006).

The longer survival time of diabetic individuals increases the chances of development of the chronic complications of the disease associated with the time of exposure to hyperglycemia. Such complications can be very debilitating to the individual and are very costly to the health system. In this context,

- cardiovascular disease is the leading cause of death in individuals with DM2;
- retinopathy represents the main cause of acquired blindness;
- nephropathy is one of the major causes of joining dialysis and transplant programs;
- Diabetic foot is an important cause of lower-limb amputations.

As a consequence of the complications of the disease, diagnostic and therapeutic procedures (catheterization, coronary bypass, retinal photocoagulation, renal transplantation, and others), hospitalizations, absenteeism, disability, and premature death can be cited, which substantially increase the direct and indirect costs of health care of the diabetic population. This

framework highlights the importance of the prevention of both the disease and its chronic complications.

The quantity of individuals with DM gives an idea of the magnitude of the problem, and estimates have been published for different regions of the world, including Brazil. Globally, 135 million had the disease in 1995, 240 million in 2005. Studies predict that this number will reach 336 million in 2030, with two-thirds living in developing countries (BARCELÓ et al., 2003; ORGANIZATION, 2011; WILD et al., 2004).

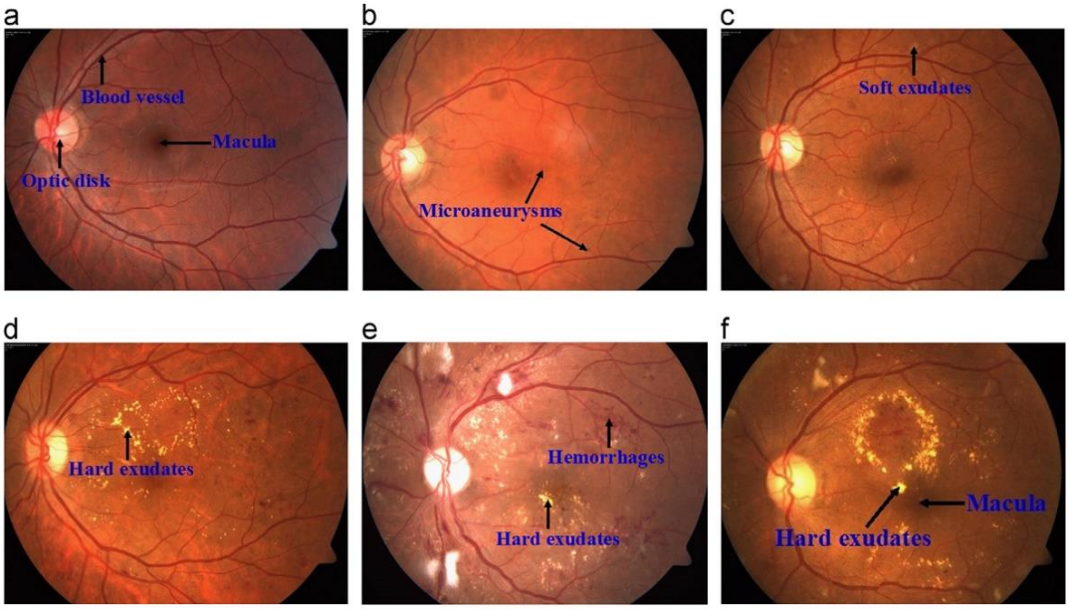
According to data from the Brazilian Ministry of Health, there are about 11 million DM patients in this country. Approximately 6.3% of the Brazilian population with 18 years or more have diabetes, equivalent to 8.3 million people. Of this group, 3 million Brazilians do not know that they have the disease (SAÚDE; ESTRATÉGICAS., 2001).

Diabetic retinopathy (DR) is caused by a failure in glycemic control in hyperglycemic patients. All diabetic patients may eventually develop DR (ALGHADYAN, 2011). The incidence rate of DR is 50% after ten years and 90% after 30 years of acquired DM. Usually, patients do not develop DR within five years from the onset of DM or before puberty. Approximately 5% of patients with DM2 have DR (HARNEY, 2006). Uncontrolled DM and its complications lead to DR which can result in vision loss and blindness. Patients with Proliferative Diabetic Retinopathy (PDR) have increased risk of heart attack, infarction, diabetic nephropathy, amputation, and death (AHMAD FADZIL et al., 2011; HARNEY, 2006).

Early stages of DR may be clinically asymptomatic, and when the disease is detected at an advanced stage treatment may become difficult (YEN; LEONG, 2008). For this reason, DR must be detected as early as possible. Figure 1 shows standard and different levels of eye fundus images with DR.

Manual diagnosis requires much effort to evaluate the images. Automatic systems, therefore, can significantly reduce time, cost, and effort. The growth of diabetes cases has led to an increase in automatic analysis tools in recent years. Besides, image processing, analysis, computer vision techniques, and increasing computer speed are being increasingly used in the medical sciences and are very relevant in modern ophthalmology.

Figure 1 - Typical eye fundus images



Typical eye fundus images: (a) Normal; (b) mild non-proliferative diabetic retinopathy; (c) moderate non-proliferative diabetic retinopathy; (d) severe non-proliferative diabetic retinopathy; (e) proliferative diabetic retinopathy; (f) Macular edema.
Source: (MOOKIAH et al., 2013).

1.1 Problem definition

1.1.1 Clinical characteristics of diabetic retinopathy

Diabetic Retinopathy can cause various abnormalities in the retina, as explained below.

1. Microaneurysms (MA) - are the first visible signs of retinal damage. Abnormal permeability and non-perfusion of retinal blood vessels cause the formation of MA. It is a red dot with a size smaller than 125 μm and with margins (WILLIAMS et al., 2004).
2. Hard exudates (HE)- are leaks of lipoproteins and proteins through abnormal vessels of the retina. They appear as small white or yellow-white deposits with sharp margins. They usually appear in clusters or rings located on the outer part of the retina (ETDRS, 1991).
3. Soft exudates or cotton wool (SE) - occur due to an obstruction in an arteriole. Reduced blood flow to the retina causes ischemia of the Retinal Nerve Fiber Layer (RNFL) that affects the axoplasmic flow and causes accumulation of axoplasmic debris in the axons of the retinal ganglion cells. The accumulation of debris appears as white "hairy" lesions in the fundus images called cotton wool spots (CHUI et al., 2009).
4. Hemorrhages (HEM) - occur due to weak capillary leaks. It is found as a red dot with irregular margin and variable density. It is usually larger than 125 μm (ETDRS, 1991).

5. Neovascularization (NV) - is the abnormal growth of new blood vessels on the inner surface of the retina. These blood vessels are weak and often bleed into glassy cavities, obscuring vision (VALLABHA et al., 2004).
6. Macular Edema (ME) - is the swelling of the retina. It is caused due to the permeability of abnormal capillaries of the retina, causing leakage of fluids and solutes around the mácula. It affects the central vision (GIANCARDO et al., 2012).

1.1.2 Stages of diabetic retinopathy

Diabetic retinopathy with MA has a 6.2% chance of progressing to Proliferative Diabetic Retinopathy (PDR) within one year. Increased number of MA is an essential early feature to evaluate the progression of DR. Pre-PDR signs include venous loops, small vessel abnormalities within the retina, and many HEM blots (SCANLON, 2010).

With the progression of ischemia, there is an increase in the likelihood of developing PDR within one year. This risk of development in one year increases from 11.3% to 54.8% from the early to the advanced stage.

New blood vessels usually grow from the arterial and venous circulation. These patients have a 25.6% to 36.9% probability of having vision loss, if not appropriately treated. Also, eyes with PDR untreated for more than two years have a 7.0% chance of having vision loss and, if not treated for more than four years, have a 20.9% chance of vision loss. Vision loss drops to 3.2% in two years of treatment and 7.4% in four years of treatment.

Patients with mild DR do not require any specific treatment other than controlling diabetes and associated risk factors such as hypertension, anemia, and renal failure. These patients need to be monitored closely; otherwise, they may progress to more advanced DR stages. Recently, it has been shown that pre-PDR can regress to background DR through optimal control of diabetes.

In the advanced stage of DR treatment is limited. In cases of marked neovascularization, several pan-retinal photocoagulation sessions may be necessary to prevent visual loss by vitreous HEM and traction retinal detachment. Inadequate laser treatment is one of the major causes of persistent neovascularization. Regression of neovascularization leaves phantom vessels or fibrous tissues. In most treated eyes, vision may remain stable once retinopathy remains constant. However, patients should be re-examined every 6 to 12 months.

Vitrectomy may prevent vision loss in patients with advanced DR. Both laser photocoagulation and vitrectomy generate an additional risk of loss of vision and are not useful

in reviewing loss of visual acuity. Intraocular steroid injections showed temporal improvements in visual sensitivity in patients with Diabetic Macular Edema (DME). However, these can cause an increase in intraocular pressure and development of cataracts (SCANLON, 2010).

1.1.3 The retinal image

An eye fundus image is a 2D projection of the 3D structure of the retina. The intensity of the image represents the amount of reflected light. The eye fund camera consists of a microscope and a camera attached to it. The optic system is similar to the indirect ophthalmoscope, which offers vertical and amplified observations of the inner surface of the eye. The camera acquires observations of the retina area at an angle of 30 to 50 degrees, with amplification of 2.5 times. The use of auxiliary lenses can increase this amplification up to 5 times (OPHTHALMIC PHOTOGRAPHERS' SOCIETY, 2015).

Color filters, fluorescein, and other types of dyes are used to perform the test. The following methods are used to perform the retinal examination for DR detection.

- Fundus photography (red-free) - the image is captured using the amount of light reflected by a specific wavelength band.
- Color Fundus Photography - the image is captured using reflected light in the RGB (Red, Green, and Blue) spectrum.
- Indocyanine and fluorescein angiography - the image formed is based on the number of photons emitted by the fluorescein or indocyanine dyes that are injected into the bloodstream.

In this work we use only Color Fundus Photography since it has a couple of advantages over the other methods described above. The red-free fundus photography uses only the green channel of the RGC fundus image, which might result in loss of valuable information. On the other hand, the fluorescein angiography requires contrast injection, which makes the exam more expensive and invasive.

1.2 Hypothesis

This work proposes an automatic DR detection system composed by a quality assessment model and a lesion localization model followed by a simple protocol to infer the DR level.

A complete computer-aided diagnostic system for DR detection must start by evaluating the quality of the retinal images since any analysis done in poor quality images (or non-retinal images) would be useless.

A recent study on a big retinal dataset called UK Biobank showed that more than 30% of retinal images acquired by professional retina acquisition devices do not have enough quality to be evaluated by an ophthalmologist (WELIKALA et al., 2016). In this context, a quality evaluation system would not only reduce errors generated by the retinal image classification algorithms but also reduce the necessity of the patients to return back to the health center to take another fundus photograph to replace the poor-quality ones.

In the following stage, two steps compose the automatic DR detection system: i) lesion localization and ii) DR detection using retinal images.

The lesion localization step is not strictly required, especially for systems based on deep learning since the entire image can be used as input without the classical lesion candidate selection and feature extraction. However, such models classify the images without marking the regions responsible for an eventual DR classification. Furthermore, for a real Computer-Assisted Diagnostic System (CADS), that is a disadvantage since the ophthalmologist might want to check the reasons for the system's classification. On the other hand, a system with lesion segmentation capacity would accomplish this requirement for obvious reasons.

As a hypothesis, we argue that it is possible to propose a pipeline based on quality assessment and red lesion localization to achieve automatic DR detection with performance similar to experts considering that a rough segmentation is sufficient to produce a discriminant marker of a lesion.

Figure 2 illustrates the proposed pipeline.

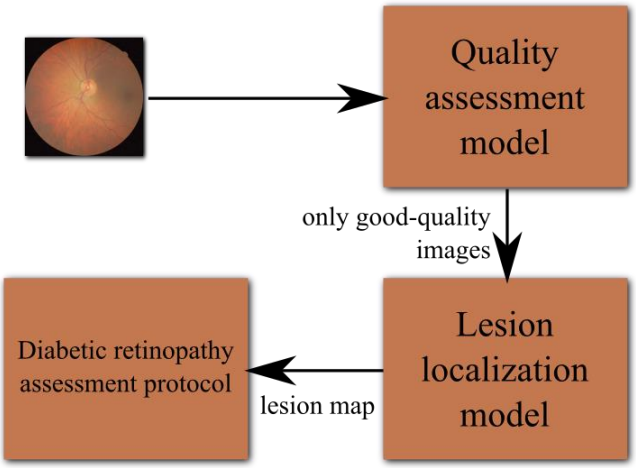
1.3 Objectives of the work

In general, the objective of this work is to propose an automatic DR detection pipeline composed by a quality assessment model, followed by lesion localization model.

More specifically, the objectives are

- To propose models that work with different datasets,
- To verify the importance of removing poor-quality images on the automatic DR detection performance,
- To reduce the number of images needed to train the lesion localization model,
- To reduce the computational cost of the lesion localization model.

Figure 2 - Proposed pipeline



Source: author's own

1.4 Contributions

A couple of contributions have been made through the two parts of this thesis. In the quality assessment chapter, a robust retinal image quality assessment model using deep learning and transfer learning is presented whose results outperform state of the art (ZAGO et al., 2018).

Next, in Chapter 3, a patch-based retinal image lesion segmentation approach (ZAGO et al., 2020) is proposed underlying the following contributions:

- Fully Convolutional Neural Networks (CNN) method for lesion localization (without any manually designed feature), acting on patches extracted from an image that we finally use to give a DR diagnosis.
- The strides usage (subsampling of patches) accelerates up to a factor of 25 the processing time compared to other CNN that does not use strides.
- During the learning stage, the sample selection method helps the final model to focus on challenging samples, which increases the performance.
- The designed model is shown robust to cross-validation over different databases, which is promising for practical applications.

1.5 Text structure

In Chapter 2 we describe the retinal image quality assessment model, both related works, the method description, and the results are referenced in this chapter.

Next, in Chapter 3 the lesion localization approach is fully described.

Finally, Chapter 4 contains the conclusions and future works possibilities.

2 RETINAL IMAGE QUALITY ASSESSMENT

2.1 Introduction

A study involving an extensive retinal image database indicated that more than 25% of the images do not exhibit sufficient quality to allow a proper medical diagnosis (MACGILLIVRAY et al., 2015). In addition to the financial investment required to reacquire photographs of poor quality, it is inconvenient for a patient to return to a medical center to repeat the fundus photograph exam. Therefore, an automatic system for quality assessment can reduce the aforementioned problems by obtaining a second photograph immediately after the poor-quality photograph is taken.

Hence, the goal of the present study is to develop a robust system for automatic retinal image quality assessment. CNN is an a priori interesting tool for the task as observed in several other applications. It exhibits the ability to extract generic features from a large set of images. Therefore, we considered a CNN pretrained on non-medical images to extract extremely general image features. Thus, our approach is extremely different from previous studies employing CNN for retinal images.

Furthermore, following the recent trend on fine-tuning (TAJBAKHSH et al., 2016), we also adjusted the network parameters of the general network to extract retinal image- specific features by using only a small quantity of labeled images.

2.2 Related Works

Several researchers have worked on automatic retinal image quality assessment using various algorithms. The algorithms are divided into three groups based on the (i) generic image quality indicators, (ii) retinal image structures, and (iii) both generic image quality indicators and retinal structures. The main studies are reviewed in the following paragraphs and listed in Table 1 on page 35.

The algorithms that use generic image quality indicators do not segment the retinal image nor use its anatomical structure. These methods typically use simpler techniques requiring less computational power.

Lee and Wang (LEE; WANG, 1999) were the pioneers in Retinal Image Quality Assessment (RIQA) and used histogram models to create an index (Q index) to classify retinal images into two classes: good or poor quality. Lalonde et al. (LALONDE; GAGNON;

BOUCHER, 2001) modified Lee et al.'s models to operate locally instead of globally, and they graphically indicated that their features can separate good- and poor-quality retinal images in their database. Generic image features were also used in (BARTLING; WANGER; MARTIN, 2009) and evaluated on a private database by a group of specialists, who obtained a kappa agreement coefficient of $\kappa = 0.55$, which is used to measure inter-rater reliability. If two raters completely agree, then $\kappa = 1$ and if there is no agreement between them, $\kappa = 0$.

Pires et al. (PIRES DIAS; OLIVEIRA; DA SILVA CRUZ, 2014) published one of the most relevant studies on the subject. They used several public databases in conjunction with two proprietary ones. Their features were extracted according to focus, contrast, and illumination histograms, and they tested several classifiers. They reported an Area Under the Curve (AUC) of the receiver operating characteristic curve for a quality evaluation of 99.87%. Moreover, in 2014, Nugroho et al. (NUGROHO et al., 2015) used a naive Bayes classifier and contrast measurement to classify 47 images. They obtained a sensitivity (Se) of 97.6%, a specificity (Sp) of 80.0%, and an accuracy (Acc) of 89.3%.

Yao et al. (YAO et al., 2016) used generic quality parameters including entropy, texture, symmetry, and frequency components, and a Support Vector Machine (SVM) classifier to obtain an Sp of 93.08% and an AUC of 96.19%.

In (MAHAPATRA et al., 2016), the authors used convolutional neural networks (CNNs) to assess the retinal image quality. They combined features from a local saliency map and from a CNN trained on more than 20,000 images of a proprietary database using random forest and SVM classifiers. After the cross-validation, they obtained a Se of 97.9%, a Sp of 97.8%, and an Acc of 97.9%.

Abdel-Hamid et al. (ABDEL HAMID et al., 2016) proposed a classifier based on a wavelet and a hue-saturation-value space to assess the sharpness, illumination, homogeneity, Field Of View (FOV), and outlier features. Their study used several databases and achieved an AUC of 99.8% via an SVM classifier.

Recently, Saha et al. (SAHA et al., 2018) proposed a methodology similar to ours. They used a pretrained CNN and a subset of the EyePACS dataset labeled by three specialists (only one of whom was an ophthalmologist) to fine-tune their network. Their algorithm was tested on a subset of the same database with 123 poor-quality and 3,302 good-quality retinal images. Although similar methods were used, in our work, novel scenarios to assess the robustness of our approach to different datasets are proposed (called inter-dataset cross-validation, where CNN is trained on a database and tested on another).

Alternatively, algorithms based on retinal image anatomical structure use specific characteristics of the retinal image to assess its quality. Typically, the methods initially segment the structures, such as the blood vessels, and subsequently use the elements to assess the image quality.

Usher et al. were the first to use segmented vessels to infer the image quality (USHER; HIMAGA; DUMSKYJ, 2003). They used a private database consisting of retinal images from 2546 patients and employed a kappa agreement coefficient to compare the results of the study to the annotation of the experts ($\kappa = 0.67$). In 2006, Niemeijer et al. (NIEMEIJER; ABRÀMOFF; VAN GINNEKEN, 2006) applied image structure clustering (ISC) to extract features and an SVM classifier to obtain an AUC of 99.68% on a proprietary database of 1,000 retinal images.

Table 1 - Review of extant studies on retinal image quality

Methodology	Year	Database	Approach	Performance
Usher et al. (USHER; HIMAGA; DUMSKYJ, 2003)	2003	Proprietary	Blood vessels segmentation, kappa index	κ: 0.67
Niemeijer et al. (NIEMEIJER; ABRÀMOFF; VAN GINNEKEN, 2006)	2006	Proprietary	Histograms, image structure clustering (ISC), SVM	AUC: 99.68%
Bartling et al. (BARTLING; WANGER; MARTIN, 2009)	2009	Proprietary	Sharpness, ilumination	κ: 0.55
Davis et al. (DAVIS et al., 2009)	2009	MESSIDOR, Proprietary	Artificial added noise, CieLAB, PLS	AUC: 99.3%

Methodology	Year	Database	Approach	Performance
Paulus et al. (PAULUS et al., 2010)	2010	Proprietary	Clustering, sharpness, texture	Se: 96.90% Sp: 80.00% Acc: 91.70% AUC: 95.30%
Hunter et al. (HUNTER et al., 2011)	2011	Proprietary	Rule-based	Se: 100.0% Sp: 93.00% Acc: 94.00%
Yu et al. (YU et al., 2012)	2012	Proprietary	Global histogram, texture, blood vessels segmentation, non-reference perceptual sharpness metric, PLS	Se: 99.00% Sp: 80.00% AUC: 98.10%
Nugroho et al. (NUGROHO et al., 2015)	2015	HEI-MED	Blood vessels segmentation, Naive Bayes	Se: 97.60% Sp: 80.00% Acc: 89.30%
Pires Dias et al. (PIRES DIAS; OLIVEIRA; DA SILVA CRUZ, 2014)	2014	DRIVE, Messidor, ROC, STARE, Proprietary	Histograms, Neural Networks	Se: 99.49% Sp: 99.76% AUC: 99.87%
Yao et al. (YAO et al., 2016)	2016	Proprietary	Statistical characteristics, entropy, texture, symmetry, frequency components,	Se: 93.08% Acc: 91.38% AUC: 96.19%

Methodology	Year	Database	Approach	Performance
			blur metric, SVM	
Welikala et al. (WELIKALA et al., 2016)	2016	UK Biobank	Blood vessels segmentation, SVM	Se: 91.59% Sp: 92.49% AUC: 98.28%
Abdel-Hamid et al. (ABDEL-HAMID et al., 2016)	2016	DRIMDB, DR1, DR2, HRF, MESSIDOR	Wavelet, sharpness, illumination, homogeneity, field definition, SVM	AUC: 94.40%
Mahapatra et al. (MAHAPATRA et al., 2016)	2016	Proprietary	CNN, SVM	Se: 97.90% Sp: 97.80% Acc: 97.90%
This work (ZAGO et al., 2018)	**2018**	**DRIMDB, ELSA-Brasil**	**Inter-databases cross-validation, CNN**	**AUC: 99.98%,** **Se: 97.10%** **Sp: 100.0%** **κ : 0.97** **(tested on DRIMDB)** **AUC: 98.56%** **Se: 92.00%** **Sp: 96.00%** **κ: 0.88** **(tested on ELSA)**

Source: author's own

Elliptical local vessel density was the technique used in (GIANCARDO et al., 2008) as a quality metric for retinal images. The authors used automatic extraction of blood vessels and classified the images from their proprietary database composed of 84 images via the k-nearest neighbor algorithm. Their proposed method achieved an accuracy of 100% on the identification

of good-quality images, 83% on that of fair-quality images, 0% on that of poor-quality images, and 66% on that of outlier images. Hunter et al. (HUNTER et al., 2011) based their study on the UK National Screening Committee guidelines and achieved an Se of 100%, a Sp of 93%, and an Acc of 94% in a database with 200 retinal images.

More recently, a novel database named UK Biobank, (UK BIOBANK, 2013) which is extremely rich in terms of size and variety of health data records, has been provided to researchers. It shows the potential to bring new pertinent results to the health care community generally. Welikala (WELIKALA et al., 2016) used blood vessel features and an SVM classifier to analyze the quality of the retinal images of UK Biobank. From a subsample of 800 images, they obtained a Se of 95.33% and a Sp of 91.13%. An important conclusion of their study is that more than 26% of the images from UK Biobank (and potentially from all the large studies that involve fundus photography) were considered as inadequate. Another study from the same group (MACGILLIVRAY et al., 2015) indicated that only 36% of the patient records had both eye images that were appropriate for analysis, thereby confirming the importance of good automatic quality assessment systems.

Finally, other studies were based on both generic and specific features.

Paulus et al. (PAULUS et al., 2010) used generic features including sharpness and texture in conjunction with unsupervised clustering to group the anatomical structures to assess the quality of 301 retinal images and obtained an AUC of 95.3%. Yu et al. (YU et al., 2012) used generic and vessel features with a Partial Least Square (PLS) classifier on a database with 1884 retinal images and obtained an AUC of 95.8%.

A previous review showed that several methods have already been proposed to detect retinal images of adequate or inadequate quality. Thus, several complex features have been introduced, each of which implies an important development stage. They are difficult to compare in terms of performance as each of them was designed and tested on a specific database for a given set of parameters including contrast and illumination. Therefore, it is difficult to predict their performance under different experimental conditions. Hence, a robust system can potentially assess the quality of images generated from different databases without any (or limited) need for adaptation. Thus, the system should rely on generic features that can represent the images present in every database.

Next, we describe the databases we used in our study and provide a full description of our approach. A set of experiments is presented to validate our solution for RIQA.

2.3 Databases

There are several publicly available databases with manual quality annotations. In this study, we selected the databases on the basis of the following criteria: (i) the database annotation was performed by experienced physicians, (ii) the images were classified as appropriate for medical analysis (gradable) or not (ungradable), and (iii) the database was public.

Therefore, the following three databases were rejected according to the above criteria:

- The HRF dataset (KÖHLER et al., 2013). This database consists of 18 image pairs from 18 human patients. Previous studies already noted that a few poor-quality images of the database only exhibited a slight blur that does not define the image as ungradable (ABDEL-HAMID et al., 2016). The study itself did not explicitly specify the criteria used to annotate the database or specify the annotation method.
- The DR1 and DR2 databases. Both databases were used in (ABDEL-HAMID et al., 2016) and (PIRES et al., 2012), and also present a few problems. The images were annotated by specialists. However, these studies did not clarify the criteria that should be used by the physicians to separate the retinal images. Additionally, after a manual inspection, it was observed that there are images with doubtful quality annotations, thereby indicating that it is necessary to review the annotations of the databases.

We finally selected two databases, namely, DRIMDB and ELSA-Brasil, that met the proposed requirements.

The Diabetic Retinopathy Image Database (DRIMDB) is a retinal image database created to evaluate the performance of quality assessment algorithms (SEVIK et al., 2014). It is composed of 216 retinal images divided into three classes: good (125), poor (69), and outlier (22 that were not used). It was created by the Retina Department of Ophthalmology, Medical Faculty, Karadeniz Technical University, Trabzon, Turkey. All images were captured by a Canon CF-60UVi fundus camera via a 60^{o} FOV and stored as JPEG files at a resolution of 570×760 pixels. The images were annotated by an expert, and good-quality images are suitable for use in medical diagnosis by an ophthalmologist.

The aim of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) is to investigate the incidence and progression of diabetes and cardiovascular diseases and their biological, behavioral, environmental, occupational, psychological, and social factors in a

cohort of adults (AQUINO et al., 2012). The study population is composed of 15,105 active or retired employees of six public institutions from different regions in Brazil. The retinal images were annotated by experienced ophthalmologists with respect to diseases and quality. Although the size of the dataset is big, only a fraction of the records in the database were available to the author. Thus, the samples used in the study were composed of 25 poor-quality images and 817 good-quality images. The retinal images were centered on the macula and optic disc of each eye and were obtained using a Canon CR-1 nonmydriatic system with an EOS 40D (10-megapixel) digital camera (Canon, Tochigiken, Japan).

2.4 Methods

2.4.1 General description of the model

The method consists of a preprocessing stage to crop the images within their Region of Interest (ROI) to reduce the quantity of data that should be processed by the network. It should be noted that no specific preprocessing is performed as a function of the different specificities of the images in the database, and, thus, the step remains extremely simple and generic.

The resulting image is presented as the input of a CNN and pretrained with non-medical images. In the last stage, the last fully connected layers are retrained on a specific DR database to increase the global performance by learning the specific problem at hand.

2.4.2 Preprocessing

To crop the retinal image around the ROI, the method[1] first blurs the image by using a 5×5 Gaussian kernel to remove any outlier pixel from the background. Subsequently, the maximum background intensity max_{bg} in the first and last $\lfloor \frac{W}{32} \rfloor$ columns of the blurred image is considered, where W corresponds to the number of columns of the image.

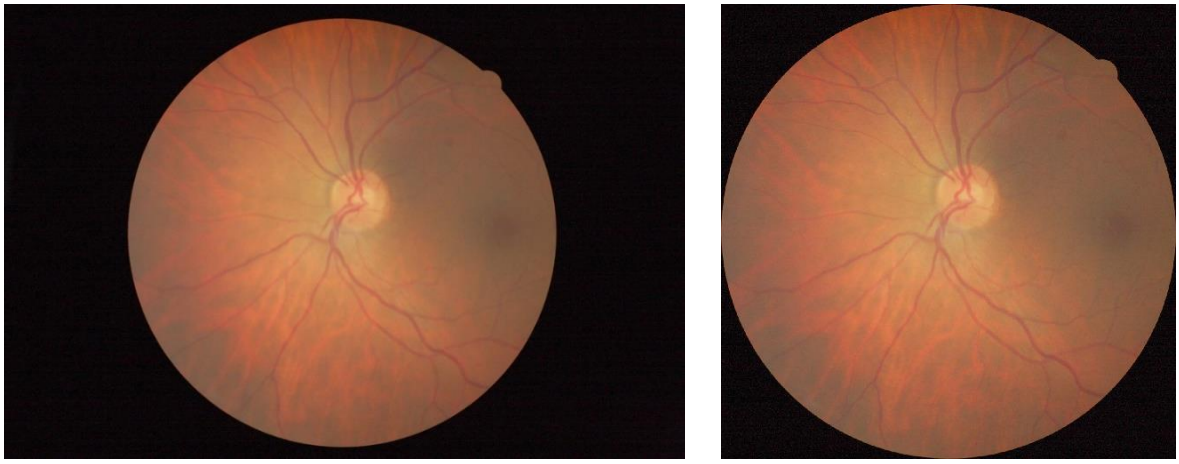
This is followed by selecting the foreground pixels by considering every pixel in the blurred image, with the intensity exceeding $max_{bg} + 10$. The image is cropped by using the smallest rectangle that contains all the foreground pixels. Figure 3 shows an example of the preprocessing stage.

Before being presented to the CNN, all RGB channels of the images are normalized from the interval [0, 255] to [−1, 1].

[1] https://github.com/sveitser/kaggle_diabetic

Figure 3 - Cropping illustration

(a) (b)



(a) Initial retinal image and (b) corresponding preprocessed retinal image.
Source: author's own

2.4.3 Data augmentation

Deep neural networks are better trained with a large amount of data. However, if the dataset exhibits a limited amount of sample images, then it is possible to create new fake samples that are added to the available data. This is known as data augmentation (GOODFELLOW et al., 2014). The new samples are obtained through simple transformations of the initial images and correspond to "real-world" acquisition situations. Therefore, the following random transformations are performed on the available images:

- Rotation: The image is randomly rotated around its center.
- Vertical and horizontal shift: The image is randomly shifted vertically and horizontally.
- Scale: The image is zoomed in or out by a factor.
- Horizontal and vertical flip.
- Contrast augmentation: Different contrast conditions are simulated by applying a transformation defined by $y = 128 + \alpha(x - 128)$, where x denotes the original image and α denotes a random number (QUELLEC et al., 2017).

The values used in the random transformations are shown in Table 2 and follow (QUELLEC et al., 2017).

Table 2 - Values used in data-augmentation transformations

Transformation	Values
Rotation	$[0^o, 360^o]$
Vertical and horizontal shift	$[0, image\ width \times 7\%]\ pixels$
Scale	[0.85,1.15]
Horizontal and vertical flip	50% flip probability
Contrast augmentation	$\alpha \in [0.6,1.67]$

Source: author's own

2.4.4 Deep neural networks

The modern term "deep learning" is related to machine learning frameworks with multiple levels of composition (GOODFELLOW et al., 2014). With respect to the subject of image processing, an extremely commonly used framework for deep learning is the CNN, which is an artificial neural network (ANN) with a special structure that is invariant to both rotation and translation. CNNs are typically composed of convolutional, pooling, and fully connected layers. An example of 2-D convolution is illustrated in Figure 4.

In convolutional layers, a new feature map is calculated by the convolution between a learned kernel and the previous feature map (or input, for the first hidden layer) followed by an activation function. Currently, rectified linear units (ReLUs) (NAIR; HINTON, 2010) correspond to the most used activation function and are defined in Equation (1):

$$f(x) = \max(0, x). \quad (1)$$

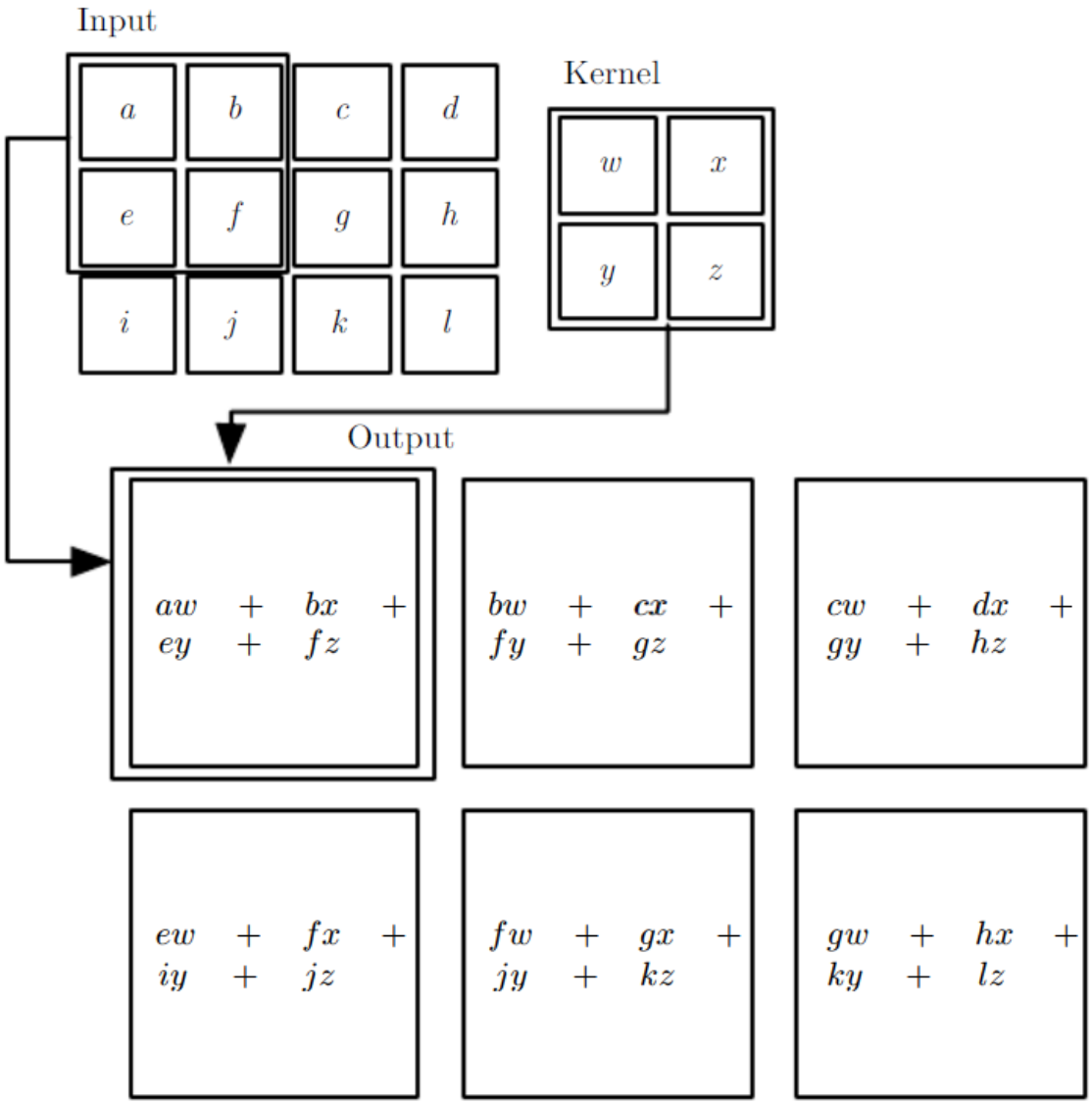
The adoption of ReLUs solves the issue of vanishing gradients, which hampers the training process of deep ANNs that use nonlinear activation functions (such as tanh or sigmoid).

Conversely, the goal of the pooling layers is to accomplish shift invariance by reducing the spatial resolution of the feature maps. This is achieved by applying a simple moving window operation (the most frequently used operations include the average, maximum, and minimum) in the previous layer output instead of using a kernel followed by an activation function as performed by the convolutional layer. Finally, fully connected layers correspond to standard ANN structures in which every output unit is connected to all units in the previous feature map (GOODFELLOW et al., 2014).

Deep CNNs exhibit an extremely high capacity of generalization because of their invariance to shift, rotation, and scale, and also because of the number of trainable parameters that they have, which can reach to the order of millions. However, the training process has its own issues. First, a large amount of labeled data is required to avoid under-fitting, and this can

be a problem for medical images. Additionally, a convergence issue exists and typically imposes several adjustments in the training parameters and architecture alike.

Figure 4 - An example of 2-D convolution



In this case we restrict the output to only positions where the kernel lies entirely within the image, called "valid" convolution.
Source: (GOODFELLOW; BENGIO; COURVILLE, 2016)

An alternative to training a CNN from scratch involves fine-tuning a network that was trained on a large unrelated labeled dataset. This technique has already been applied successfully in several areas of computer vision (AZIZPOUR et al., 2015; PENATTI; NOGUEIRA; DOS SANTOS, 2015; RAZAVIAN et al., 2014) and can reduce the aforementioned problems.

2.4.5 The architecture

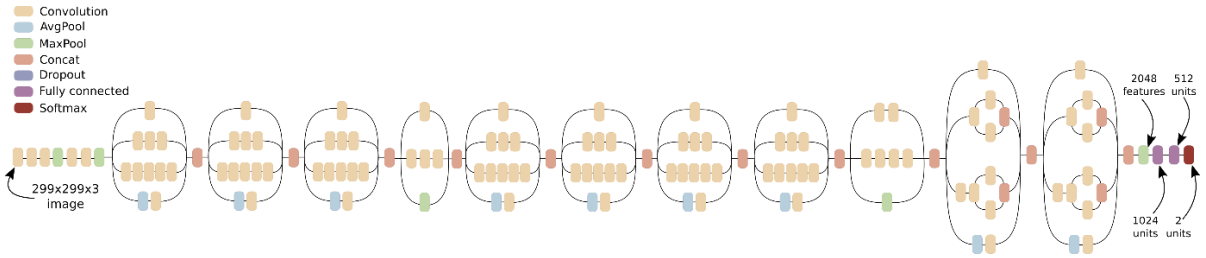
In order to use the fine-tuning technique, we must choose an architecture whose tuned weights are available. A few architectures meet this requirement and we choose Google's Inception v3 (SZEGEDY et al., 2016) based on its good performance on other image classification tasks (SZEGEDY et al., 2016).

The model's weights were trained on the ImageNet database (JIA DENG et al., 2009) as available on the used framework. In its final configuration, the network had more than 24 million parameters that were fine-tuned. The network was trained to classify images in one of the 1,000 classes of the ImageNet database (which include animals, plants, food, etc.). To adapt the model to our problem (which exhibits only two classes) and use only the general features, we replaced the last three fully connected layers with three layers with 1024, 512, and 2 units, respectively, as shown in Figure 5. The last layer exhibited the softmax activation function since it was desirable for the output to sum to 1, whereas the others were activated by a ReLU function. The sofmax activation function is the application of the standard exponential function to each input of the input vector and normalization of these values by dividing by the sum of all these exponentials. This ensures that the sum of the components of the output vector is 1, as describes below:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}. \qquad (2)$$

All colored images were redimensioned to $299 \times 299 \times 3$ since this dimension corresponded to the input size of Inception v3, and the output feature layer (the last one before the fully connected layers) exhibits 2048 units.

Figure 5 - CNN architecture used: Inception v3 with different final layers



Each rectangle represents a layer following according to its color.
Source: author's own

The model was trained by using the stochastic gradient descent algorithm with the following:

- a mean squared error loss function,

- a learning rate starting at 0.01 and reduced by a factor of 10 when the validation error reached a plateau,
- a momentum corresponding to 0.9,
- batches of 16 images,
- 100 steps per epoch, and
- a limit of 500 epochs or fewer if the validation loss did not improve for 10 epochs.

Using the two eye databases previously mentioned, we performed training and testing using two different protocols: (i) intra-dataset cross-validation in which three-fold cross-validation was applied by using images from the same database, and (ii) inter-dataset cross-validation in which the algorithm was trained by using images from one database and tested on another.

Table 3 - Performance of the deep neural network with different database configurations

Cross-validation	**Database**	**Fine-tuning**	**Cohen-Kappa**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**AUC (%)**
Inter-databases (mean and 95% CI)	DRIMDB	no	0.03 [0.01 – 0.04]	51.28 [47.11 – 55.45]	51.67 [47.50 – 55.84]	51.45 [47.28 – 55.62]	53.9 [50.5 – 59.7]
		yes	0.97 [0.96 – 0.99]	100.0 [100.0 – 100.0]	97.18 [95.8 – 98.56]	98.55 [97.55 – 99.55]	99.9 [99.3 – 100]
	ELSA-Brasil	no	0.60 [0.53 – 0.67]	80.00 [74.46 – 85.54]	80.00 [74.46 – 85.54]	80.00 [74.46 – 85.54]	84.4 [73.0 – 93.7]
		yes	0.88 [0.83 – 0.93]	95.83 [93.06 – 98.6]	92.31 [88.61 – 96.0]	94.00 [90.71 – 97.29]	97.1 [92.2 – 100]
Intra-databases (mean and std)	DRIMDB	no	0.91 ± 0.04	92.75 ± 2.51	98.55 ± 2.51	95.65 ± 2.17	98.74 ± 1.72
		yes	0.94 ± 0.03	95.65 ± 0.00	98.55 ± 2.51	97.10 ± 1.26	99.81 ± 0.33
	ELSA-Brasil	no	0.41 ± 0.20	79.63 ± 14.85	61.11 ± 24.06	70.37 ± 10.17	70.13 ± 23.33
		yes	0.75 ± 0.13	75.46 ± 13.20	100.00 ± 0.00	87.73 ± 6.60	97.40 ± 3.25

Source: author's own

For each experiment, 20% of the training set was used for the validation to reduce over-fitting by early stopping the training algorithm. In all cases, both the validation and the test sets were balanced, equalizing the number of samples in each class through discarding some fair-quality images. Finally, every network was trained in three ways as follows:

1. Pretrained without fine-tuning: only the three newly added layers were trained.
2. Pretrained with fine-tuning: all layers of the network were trained.
3. From scratch: the network was fully trained and was initialized with random parameters rather than with pretrained parameters.

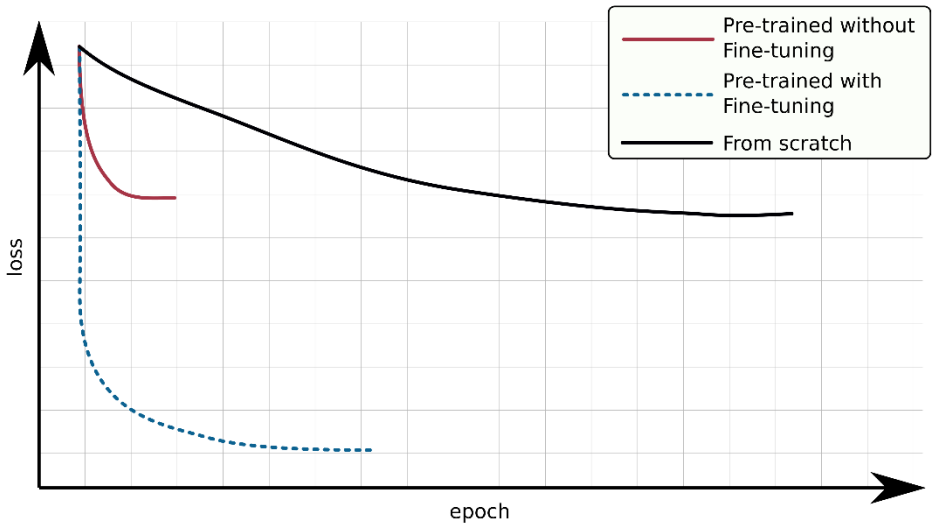
2.5 Results

The experiments were divided into the following two parts:

- Intra-database cross-validation: This involves three-fold cross-validation with images from a single database. The goal is to use the same methodology as that used in other studies for comparison purposes.
- Inter-database cross-validation: This involves training with images from a database and testing on images from another database. The experiment was performed to demonstrate the robustness of our approach to changes in the characteristics of the images, and this was an aim while designing the novel system.

In each experiment, the initial features were learned from the non-generic problem of ImageNet. Therefore, they were not specialized to the DR problem. The aim of the following stage was to specialize the network to the task at hand by using a DR database (DRIMDB and ELSA-Brasil). This was performed in two ways as follows: (i) with fine-tuning, where all weights of the model were retrained, and (ii) without fine-tuning, where only the last three layers were trained.

Figure 6 - Typical loss value course during the training process for different types of experiments



Source: author's own

The processing was performed using Keras framework using a computer with an Intel Core i7-7700 CPU (@3.60 GHz × 8), 16 GB of RAM, and a GeForce GTX 1050 Ti GPU. A time period ranging from 15 to 120 min (15 to 114 epochs) was used in training the networks

based on the method (i.e., whether fine-tuning was applied or not) and on the database used. A typical loss value course during the training process is shown in Figure 6. It should be noted that it takes more time to train a network from scratch than to fine-tune a pretrained network. Furthermore, it is faster to train only the last layers since significantly fewer parameters are fine-tuned.

Previous studies indicated that it is better to use pretrained CNNs for medical images than to train them from scratch (TAJBAKHSH et al., 2016). However, we tested the latter option to ensure that this also works for retinal images and, at least, for quality assessment purposes. The result is shown in Figure 7 and Table 3. When the network was trained from scratch, the result corresponded to an AUC of $95.02\% \pm 2.15$ for DRIMDB and an AUC of $70.06\% \pm 14.51$ for ELSA-Brasil, and this was significantly worse than that of the other training configurations. This confirms the results obtained in (TAJBAKHSH et al., 2016), wherein fine-tuning a pretrained network was better than training it from scratch. This happens because the model learns good generic features when trained with such amount of distinct images, specially in the first layers as shown in (TAJBAKHSH et al., 2016). Even if the same amount of retinal images with quality labels were available, initializing the model with pre-trained weights would be beneficial both in training speed and performance (TAJBAKHSH et al., 2016).

An initial interesting result observed from our experiments was that acceptable results were noted while training with DRIMDB in pretrained networks without fine-tuning (i.e., only the last three layers were trained, thereby indicating that extremely general features were extracted, namely, the ones used to distinguish between the ImageNet classes). This indicates that the general features extracted from the DRIMDB images are better descriptors for RIQA since they resulted in good separation for both DRIMDB itself and ELSA-Brasil (for inter-database cross-validation).

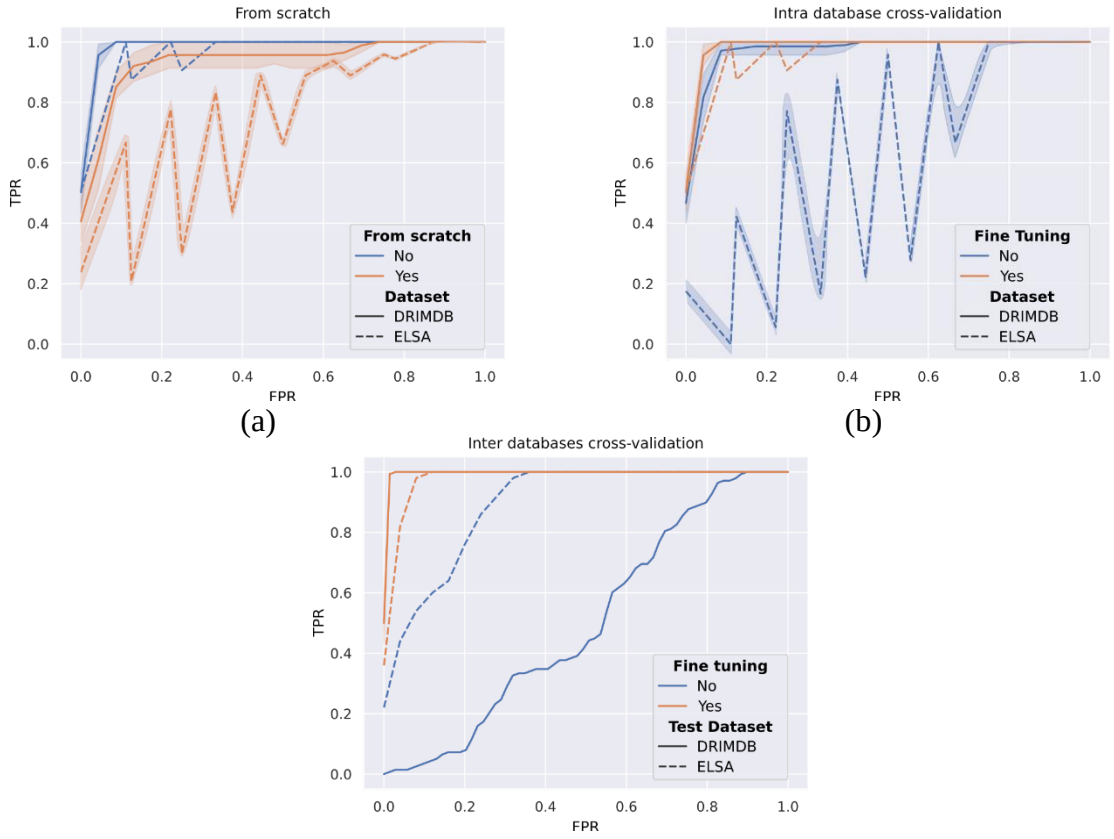
As expected, the use of fine-tuning produced better results for all databases in terms of the AUC mean and standard deviation.

The results involving fine-tuning were good when compared to those in other studies, as shown in Table 1, with an AUC of $99.81\% \pm 0.33$ for DRIMDB and an AUC of $97.39\% \pm 3.25$ for ELSA-Brasil. The improvements in the mean and, most importantly, in the standard deviation indicated that the fine-tuned network extracted better and more trustful features.

Most studies published in the subject used the same database to train and test their algorithms. This means that there is no guarantee that their approach will work on new images

because images from the same database are similar in most cases in terms of features including the color distribution and contrast distribution.

Figure 7 - Receiver operating characteristic curves for quality assessment experiments



ROC curves for networks trained (a) from scratch, (b) intra-database cross- validation, and (c) inter-database cross-validation.
Source: author's own

The experimental results indicate that the proposed method is robust with respect to the changes in a database and that the model can be used to classify retinal images in a real-world scenario since its performance was significantly high in terms of the aforementioned limited-size datasets. Specifically, we achieved an AUC of 99.98% for DRIMDB and an AUC of 98.56% via fine-tuning for ELSA-Brasil for the inter-database experiment.

An extant study (PIRES DIAS; OLIVEIRA; DA SILVA CRUZ, 2014) indicated comparable results and also used images from several databases. Nevertheless, in its experiment, all images from each class were attributed to a single database (all images from the gradable class corresponded to the MESSIDOR database, whereas all images from the ungradable class corresponded to the study's proprietary database), and, thus, it was not clear whether the classifier was learning to separate gradable/ungradable images or MESSIDOR/proprietary database images.

2.6 Conclusion

In this study, a pretrained deep neural network was adapted to deal with retinal image quality assessment. Fine-tuning was demonstrated as an efficient tool for CNN adaptation. Furthermore, the use of pretrained features avoided the burden of requiring a large amount of specific data and reduced the time spent during the learning phase. The major limitations of this work are the small number of available retinal datasets with quality labels and the small number of both fair and, mainly, poor-quality retinal images in the databases used. However, the results hold significant promise in terms of an efficient operational system specifically when the acquisition involves mobile devices. Finally, this study constitutes a first step toward the design of a global diabetic retinopathy detection system aiming in particular at an operational system for clinical usage.

3 DIABETIC RETINOPATHY DETECTION USING RED LESION LOCALIZATION AND CONVOLUTIONAL NEURAL NETWORKS

3.1 Introduction

Diabetic retinopathy (DR) is a common complication of diabetes and involves variations in retinal blood vessels. These variations can cause the blood vessels to bleed or leak fluid, distorting vision. It is the most prevalent cause of vision loss among individuals with diabetes and a significant cause of blindness among working-age adults. Fortunately, early detection, timely therapy, and adequate diabetic eye disease follow-up care can safeguard against the loss of vision.

Therefore, it is crucial to offer easy methods of detection of this disease on a large scale. Several devices enable the acquisition of retinal images, but manual diagnosis and evaluation of images requires significant effort. Thus, automatic systems can reduce time, cost, and effort significantly and be a valuable tool for practitioners, especially considering the increasing number of diabetes cases.

The first attempts at automatic systems involved the use of classical image processing techniques, but quite recently, the introduction of deep networks and, in particular, convolutional neural networks (CNNs) has had a significant effect on medical image analysis. Indeed, such approaches are producing impressive results in the classification of many types of diseases, as well as in the localization and segmentation of regions of interest (LITJENS et al., 2017). These results are obtained because of the availability of a vast quantity of labeled data. DR detection, which is the subject of the present study, is no exception, as it is generally performed through the analysis of retinal images.

In contrast to classical image processing systems, which use predefined features as an intermediate stage for classification, CNN models can directly propose a classification from raw pixel images and independently extract the appropriate representation of the images.

Following this line of thinking, several authors have proposed CNN models for indicating the degree of DR or the presence of a disease in a given image. These models are very effective, but, if learned from scratch, they require a large quantity of labeled data, which are not always available. Moreover, the performance is very dependent on the statistics of the data used for learning. Therefore, the resulting system may not be robust when used with a different type of data from a different acquisition environment.

In practice, these models, which do not explicitly explain their method of decision-making, are not suited for interaction with clinicians. Nevertheless, it is important for clinicians to understand why the model made its decision and in particular, which region in the images influenced the DR diagnosis.

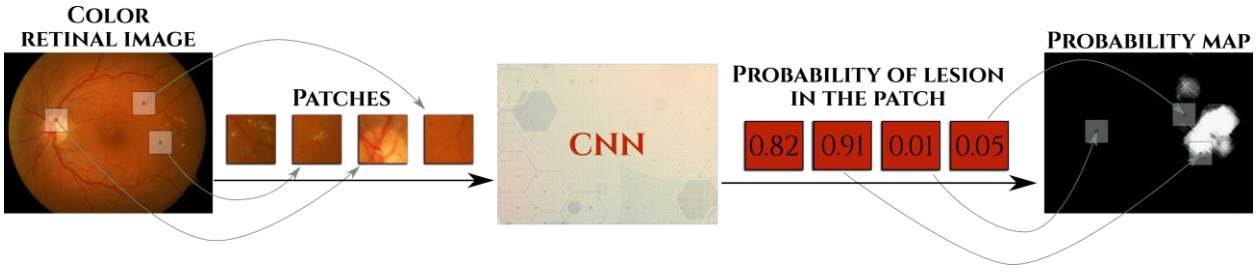
For this reason, we decided to focus the present work on building an automatic DR system based on a CNN approach, but specifically, one that explicitly relies on a localization of lesions that can be assessed by an expert. We do not aim to precisely segment the lesions, but to localize them, and to produce some probability that the lesion may induce DR or not.

The chosen method is inspired by the work of Ciresan et al. (CIREŞAN et al., 2013) in cancerology, and it uses a deep CNN to classify patches of an image as lesion or non-lesion. Therefore, the system is able to localize the regions with lesions for further DR detection.

In practice, the CNN takes patches of the initial retinal image as inputs and produces a probability of a lesion being present in each patch. The resulting probability map is post-processed to ultimately decide whether DR is present or not. This model can be seen in Figure 8.

As the first signs of DR in the retina are generally microaneurysms (MAs) and hemorrhages, we focus on localizing these lesions, called "red lesions," instead of bright lesions or neovascularizations (which are later signs of DR).

Figure 8 - Localization model based on patches



The convolutional neural network (CNN) input is a single patch of the retinal image, and, for each patch, the output is the probability of a lesion being in the analyzed patch.
Source: author's own

Training such a CNN requires the availability of a database with explicit labeling of the regions by an expert. In contrast to what occurs when a CNN is trained on a global image, even if the training database is of limited size, we obtain a large number of input data (as one image of size 512 pixels x 512 pixels gives rise to 1,310,720 patches). Therefore, finding efficient protocols for choosing adequate samples has a significant impact on the final quality of the system.

We propose a novel strategy for selecting challenging samples by relying on a two-stage construction: a simple classifier allows for the detection of misclassified samples, which are used to enrich the initial training database for the final classifier in a second stage. In this way, the performance of the classifier is increased. This is one of the solutions which makes this study original.

The high number of patches also raises a problem in terms of complexity in the production phase (once the learning phase is completed). The modification that we introduced relies on a subsampling of the image. Indeed, as we search for only a raw localization of the lesion and not a precise segmentation, we can downsample the image, and consider only a sampling of the patches for analysis. The use of strides accelerates the process by up to a factor of 25 as compared to other CNNs that do not use strides.

We also expect such an approach to be more generic than global models working globally on an image, as the final decision will result from a focused local analysis. We will show in the experimental part of the study that this is indeed the case, in particular, as compared to other state-of-the-art models.

As no large databases of labeled retinal lesions are available, we propose, like other authors, to use a pre-trained CNN with transfer learning to classify patches of retinal images as lesion or non-lesion.

3.1.1 Related Works

State-of-the-art automatic DR detection techniques can be roughly classified into two types: older studies that rely on classical image processing techniques for detecting, segmenting, and analyzing lesions in images depending on their precise characteristics, and more recent studies, that rely on CNNs to perform both feature extraction and classification.

Because MAs and hemorrhages are generally the first signs of DR, several works have focused on these diseases, particularly for the early detection of DR. Hence, we focus our review on those works.

There are many studies that use the classical image processing pipeline, and some of them are summarized in Table 4. We have limited ourselves to the studies that detect MAs and hemorrhage lesions, as their presence is significant for the early detection of DR.

Most works on this subject use the following pipeline.

- Preprocessing: enhancing the retinal image to make the lesions more visible.
- Lesion candidate selection: filtering, morphological transformations, and thresholding are the techniques commonly used to select candidates.
- Feature extraction: features such as perimeter, area, and average intensity are obtained from the lesion candidates. Commonly-used features include perimeter, area, and average intensity.
- Classification: the candidates are classified into lesion or non-lesion based on the characteristics of the extracted features.

Table 4 - Works that employ a classical image processing pipeline with different database configurations for lesion detection

Authors	**Year**	**Approach**	**Databases**	**Results**
Larsen et al. (LARSEN et al., 2003)	2003	Commercial system against experts	Welsh Community Diabetic Retinopathy Study (WCSDR)	Image level lesion detection Specificity $(Sp) = 0.714$ Sensitivity $(Se) = 0.967$ Area under the receiver's operating characteristic curve $(AUC) = 0.903$ $\kappa = 0.659$
Niemeijer et al. (NIEMEIJER et al., 2005)	2005	Classical image processing, k-nearest neighbor (k-NN)	140 images. Screening program in the Netherlands and another private one	Image level red lesion detection $Se = 1.0$ $Sp = 0.87$
Balasubramanian et al. (BALASUBRAMANIAN; PRADHAN; CHANDRASEKARAN, 2008)	2008	Automatic seed generation (ASG), implicitly hybrid classifier called spatiotemporal feature map classifier (STFM)	Private dataset of 63 images	Lesion level $Se = 0.87$ $Sp = 0.955$
Zhang et al. (ZHANG et al., 2009)	2009	Multiscale Correlation Filtering and dynamic thresholding, Coarse and fine level	Retinopathy Online Challenge (ROC) dataset	Lesion level Free-response receiver operating characteristic (FROC) curves

Authors	Year	Approach	Databases	Results
Kande et al. (KANDE et al., 2009)	2009	Matched filtering relative entropy-based thresholding morphological top-hat transformation support vector machines (SVMs)	Private dataset with 80 retinal images, Standard Diabetic Retinopathy Database, Calibration Level 0 (DIARETDB0) and Standard Diabetic Retinopathy Database, Calibration Level 1 (DIARETDB1)	Lesion level $Se = 0.962$ $Sp = 0.995$
Kande et al. (KANDE; SAVITHRI; SUBBAIAH, 2010)	2010	Histogram matching, contrast stretching, median filtering, matched filter, relative entropy-based thresholding, morphological top-hat transformation, a connected component analysis was applied to binary objects, SVMs	Private dataset, Structured Analysis of the Retina (STARE), DIARETDB0, and DIARETDB1	Image level lesion detection $Se = 1.0$ $Sp = 0.91$ $AUC = 0.962$
Zhang et al. (ZHANG et al., 2010)	2010	An approach based on multi-scale correlation filtering (MSCF) and dynamic thresholding	100 images from ROC and DIARETDB1	Lesion level FROC curves
Quellec et al. (QUELLEC et al., 2011)	2011	Filters with trained parameters, k-NN	26 images from a program in the Netherlands, 74 from Abràmoff's retinal referral clinic (private)	Diabetic retinopathy (DR) detection $AUC = 0.927$
Sánchez et al. (SÁNCHEZ et al., 2011)	2011	Testing an existing system against experts	Messidor	DR detection Expert A $AUC = 0.922$ $95\%CI = [0.902–0.936]$ Expert B $AUC = 0.865$ $95\%CI = [0.789–0.925]$ CAD System

Authors	Year	Approach	Databases	Results
				$AUC = 0.876$ $95\%CI = [0.856–0.895]$
Rocha et al. (ROCHA et al., 2012)	2012	Using a visual word dictionary to build a specific projection space for each class of interest, SVM	DR1	White lesion detection $AUC = 0.953$ red lesion detection $AUC = 0.933$
Antal et al. (ANTAL; HAJDU, 2012)	2012	Using a fusion of several preprocessing and feature extractors, simulated annealing	199 images selected from three databases, ROC, DIARETDB1, private	
Zhang et al. (ZHANG et al., 2014)	2014	Mathematical morphology, contextual features, random forest	DIARETDB1, e-ophtha EX, Messidor, Hamilton Eye Institute Macular Edema Dataset (HEI-MED)	DR Detection Trained on Messidor and tested on HEI-MED $AUC = 0.82$ Messidor $AUC = 0.93$ HEI-MED $AUC = 0.94$
Seoud et al. (SEOUD et al., 2016)	2016	Dynamic Shape Features, random forest	Six datasets	DR Detection Messidor $AUC = 0.899$

Source: author's own

The second period corresponds to the wider, more recent one, which started in the domain of DR detection, with the use of CNN-based systems. Most of the studies use CNNs globally on the images without any detection of regions (ABRÀMOFF et al., 2016; GARGEYA; LENG, 2017; GULSHAN et al., 2016b; QUELLEC et al., 2017; TING et al., 2017), and only few approaches consider lesion detection before DR classification (CHUDZIK et al., 2018; ORLANDO et al., 2018). To that end, the retinal image is split into patches that constitute the input to the network.

In that regard, a few papers have been published on lesion detection using deep learning (CHUDZIK et al., 2018; ORLANDO et al., 2018). The studies that employed CNNs are summarized in Table 5.

In particular, two studies are related to the present one, as they also employ patch-based CNN.

Orlando et al. (ORLANDO et al., 2018) selected lesion candidates using traditional image processing, and used manually-designed features together with CNN features for classification by a random forest classifier.

A lesion probability map is built by assembling the outputs of the classifier in the position of each candidate analyzed. The authors conducted several experiments, but two of them are more important in the context of this study: one experiment was conducted to evaluate the lesion localization and another to assess the DR detection capability of the proposed pipeline. In contrast to (ORLANDO et al., 2018), our approach is entirely based on deep learning, which increases the genericity concerning different datasets.

A patch-based CNN MA detection method was developed by Chudzik et al. (CHUDZIK et al., 2018). They aimed to segment the regions rather precisely to detect MA. Their processing pipeline was simple and composed of preprocessing, patch extraction, and classification. A 24-layer CNN was used to segment the potential lesions, and a voting method was employed to generate the final probability map. A Dice coefficient loss function was used to solve the problem of imbalanced data. Our work differs from (CHUDZIK et al., 2018) in several aspects. First, our goal is to localize the lesions for further DR detection instead of segmenting them, and this allows us to select only some patches of the image, thereby speeding up the prediction process. We believe that a rough segmentation is sufficient to produce a discriminant marker of a lesion. In addition, our approach does not require any retraining when the model is applied to a new dataset. The model learns to detect lesions (localization) on a given database and is tested in terms of DR on other databases.

To summarize, the main contributions and novelties brought by the present study are the following.

- We designed a fully automatic CNN method for lesion localization (without any manually-designed features), acting on patches extracted from an image that are ultimately used to provide a DR diagnosis.

- The usage of strides (subsampling of patches) accelerates the processing time by up to a factor of 25 as compared to other CNN methods that do not use strides.
- During the learning stage, the sample selection method helps the final model focus on challenging samples, thereby increasing the performance.
- The designed model is shown to be robust to cross-validation over different databases, which is promising for practical applications.

Section 3.2.1 is devoted to a description of the databases considered in this work and a precise explanation of the main stages of the model. It includes a description of the CNN, procedure for patch labeling, challenging patches selection for the training, subsampling method for accelerating the lesion localization, and process of DR detection from the global image. In Section 3.3, we present our results by using cross-dataset validation to prove the robustness of our method and compare them with those from recent literature.

Table 5 - Summary of the papers that employed deep learning for DR or lesion detection

Authors	**Year**	**Approach**	**Databases**	**Results**
Abràmoff et al. (ABRÀMOFF et al., 2016)	2016	Set of image-level convolutional neural networks (CNNs) followed by random forest	Trained on up to 1,250,000 images, manually annotated, tested on Messidor2	Need for referral DR $Se = 0.968$ $95\%CI = [0.933 - 0.988]$ $Sp = 0.870$ $95\%CI = [0.842 - 0.894]$ $AUC = 0.980$ $95\%CI = [0.968 - 0.992]$
	2016	An ensemble of InceptionV3 networks,	Private, Eye Picture Archive	Need for referral DR on Messidor-2

Authors	Year	Approach	Databases	Results
Gulshan et al. (GULSHAN et al., 2016b)		trained on 100,000+ images, multiple grades per image, two images per subject	Communication System (EyePACS)-1, and Messidor-2	$AUC = 0.999$ $95\%CI = [0.986 - 0.995]$
Gargeya et al. (GARGEYA; LENG, 2017)	2017	Small image-level CNN, data augmentation using rotation, brightness, and contrast, trained on 75,000+ images	EyePACS-1 (train) and Messidor-2 (test)	Need for referral DR on Messidor-2 $AUC = 0.94$ $Se = 0.93$ $Sp = 0.87$
Ting et al. (TING et al., 2017)	2017	Ensembles of eight image-level CNNs, all using an adaptation of the Visual Geometry Group (VGG)Net architecture	Private dataset	DR Screening $AUC = 0.936$ $95\%CI = [0.925 - 0.943]$ Need for referral $AUC = 0.958$ $95\%CI[0.956 - 0.961]$
Quellec et al. (QUELLEC et al., 2017)	2017	Image-level CNN with pixel-level visualization, an ensemble of CNNs	Private, Kaggle, DIARETDB1	Need for referral on Kaggle test-set $AUC = 0.954$
Chudzik et al. (CHUDZIK et al., 2018)	2018	Patch-based CNN, dice loss function.	E-Ophtha, DIARETDB1, and ROC	FROC Curves
Orlando et al. (ORLANDO et al., 2018)	2018	Classical candidate selection, patch-based CNN followed by random forest	E-Ophtha, DIARETDB1, and Messidor	DR Screening on Messidor $AUC = 0.893$ $Se = 0.911$

Authors	Year	Approach	Databases	Results
		with manually designed features added		Need for referral on Messidor $AUC = 0.934$ $Se = 0.972$ $AUC = 0.954$

Source: author's own

3.2 Materials and Methods

3.2.1 Databases

Seven datasets were used in our experiments: the Standard Diabetic Retinopathy Database. Calibration level 0 (DIARETDB0), Standard Diabetic Retinopathy Database. Calibration level 1 (DIARETDB1) (KAUPPI et al., 2007), Kaggle, Messidor (DECENCIÈRE et al., 2014), Messidor-2, Indian Diabetic Retinopathy Image Dataset (IDRiD) (PORWAL et al., 2018b), and DDR (LI et al., 2019).

DIARETDB0 (KAUPPI et al., 2006) consists of 130 fundus images; 20 are normal, and 110 contain signs of DR. The dataset has annotations concerning the presence of red small dots, hemorrhages, hard exudates, soft exudates, and neovascularization on the image level.

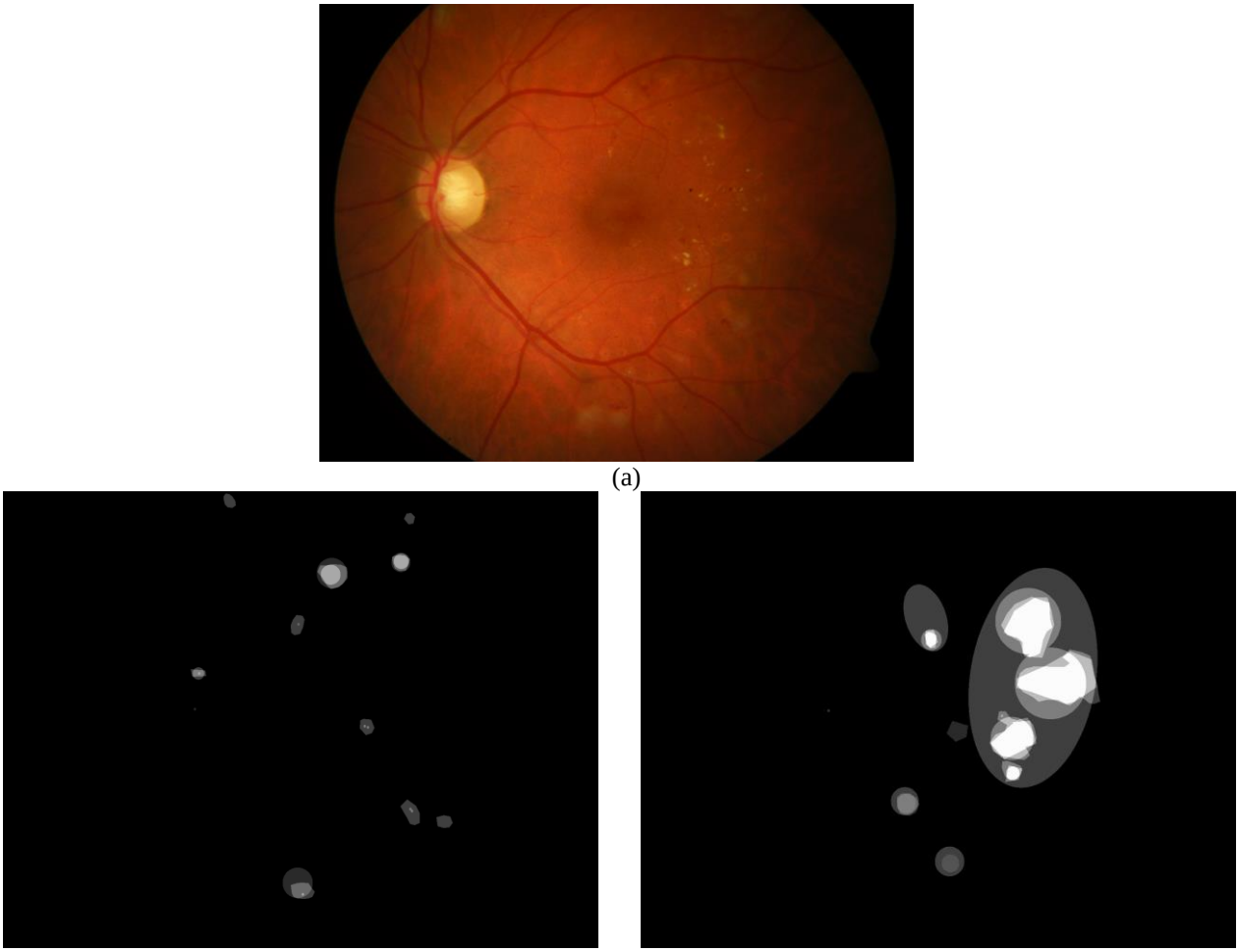
DIARETDB1 (KAUPPI et al., 2007) is composed of 84 retinal images with signs of DR, and 5 considered normal. It contains 28 images in the training set, and 61 in the test set. The entire dataset was analyzed by four experts who delineated the lesions by type: soft exudates, hard exudates, red small dots, and hemorrhages. Five levels of lesion agreements are possible, 0, 0.25, 0.5, 0.75, and 1, indicating how many experts labeled each pixel as a lesion. Figure 9 shows a retinal image and its ground truths from this dataset. The different gray levels in the ground-truth images indicate the level of agreement. Each pixel is associated with one of 5 values (0 means no lesion, whereas the others indicate a lesion with a certain degree of confidence).

The Kaggle retinal dataset is available in a competition platform website and was proposed for a DR detection competition in 2015[2]. It consists of 88,702 color fundus images, including 35,126 for training and 53,576 for testing. There are two images per subject, one per

2 https://www.kaggle.com/c/diabetic-retinopathy-detection/

eye. An expert evaluated each image for the presence of DR with a scale of 0–4, according to the Early Treatment Diabetic Retinopathy Study (ETDRS) scale (WILKINSON et al., 2003).

Figure 9 - Retinal image and its lesions ground truths



(a) (b) (c)

For each type of evaluated lesion, the Standard Diabetic Retinopathy Database Calibration Level 1 (DIARETDB1) provides a ground truth image composed by the superposition of each expert's annotation. (a) shows retinal image and ground truths, (b) shows the hemorrhages' ground truth, and (c) shows the red small dots' ground truth. The brighter the dots appear, the more the experts agree. Source: (KAUPPI et al., 2007)

Messidor contains only image-level labels and indicates DR presence through macular edema grades. It comprises 1,200 fundus images acquired from different ophthalmic institutions in France (DECENCIÈRE et al., 2014). There are four levels of DR in this dataset: R0, composed of images indicating a healthy retinal image; R1, composed of images with very mild signs of DR; R2, composed of images with signs of DR that require the attention of an ophthalmologist; and R3, composed of images indicating a proliferative DR retinal image.

The public Messidor-2 dataset contains 1,748 fundus images from 874 subjects in a region of France. Although the publishers do not provide public labels, referable DR annotations are available from another research group[3].

Another dataset used in this work is IDRiD, which contains 516 images annotated according to a DR severity level (0-4). The dataset is divided into a training (413) and test set (103).

The last dataset used in the present work is the DDR dataset, acquired from 147 hospitals in China. It consists of 13,673 fundus images, annotated with the severity of DR as determined by multiple experts. The test set is composed of 4,105 images. These datasets are summarized in Table 6.

In our work, DIARETDB1 was used to train the models, as it labels lesions at the pixel level. In that regard, all of the remaining datasets were used to validate the models for DR classification, as they provide DR-level annotations. When possible, a comparison with other approaches in the literature on the same dataset is provided.

3.2.2 Preprocessing

The retinal images (I_{ret}) are enhanced using a high-boost filter as proposed in (VAN GRINSVEN et al., 2016), to make the lesions more visible:

$$I_{en}(\sigma) = \alpha \cdot I_{ret} + \tau \cdot Gaussian(0, \sigma) * I_{ret} + \gamma. \quad (3)$$

Here, $\alpha = 4$, $\tau = -4$, $\sigma = \chi/30$, $\gamma = 128$, χ is the width of the image, I_{ret} values are between 0 and 255, and the operator $*$ represents a 2D convolution. This pre-processing tends to remove the background and enhance any structure with color variations such as lesions and blood vessels. The result of the preprocessing can be observed in Figure 10. This preprocessing is performed on all datasets.

3.2.3 Model Training

3.2.3.1 Patch labeling

The patch-based method is inspired by Cireşan et al.'s work (CIREŞAN et al., 2013) on cancer detection. A patch is a $p \times p$ region centered on a particular pixel.

---

3 https://medicine.uiowa.edu/eye/abramoff

Table 6 - Description of the datasets used in this work

Dataset	Number of images/subjects	No DR (%)	DR at any level	Camera	Number of referees
DIARETDB0 (KAUPPI et al., 2007)	130	20 (15.4)	110 (84.6)	Unknown	not informed
DIARETDB1 (KAUPPI et al., 2007)	89	5 (5.6)	84 (94.4)	Unknown	4
Kaggle training set[* 4]	15,919	11,583 (73.4)	4,186 (26.5)	Various	1
Messidor (DECENCIÈRE et al., 2014)	1,200	546 (45.5)	654 (54.5)	Topcon	not informed
Messidor-2[* 5]	874	684 (78.2)	190 (21.7)	Topcon	3
Indian Diabetic Retinopathy Image Dataset (IDRiD) test set (PORWAL et al., 2018a)	103	34 (33)	69 (67)	Kowa	not informed
DDR (LI et al., 2019)	4,105	1,880 (45.8)	2,225 (54.2)	Topcon	4

[*] Two images per subject

Source: author's own

We aim to provide a classification of each patch extracted from a given retinal image into two clusters: lesion or non-lesion. (We could imagine more classes, as DIARETDB1 provides finer labeling, but we want to make the entire process as simple as possible.)

As DIARETDB1 provides labels for several types of lesions, the first step is to combine the labels of the types of lesions we are interested in – hemorrhages (I_{label}^{hem}), and red small dots

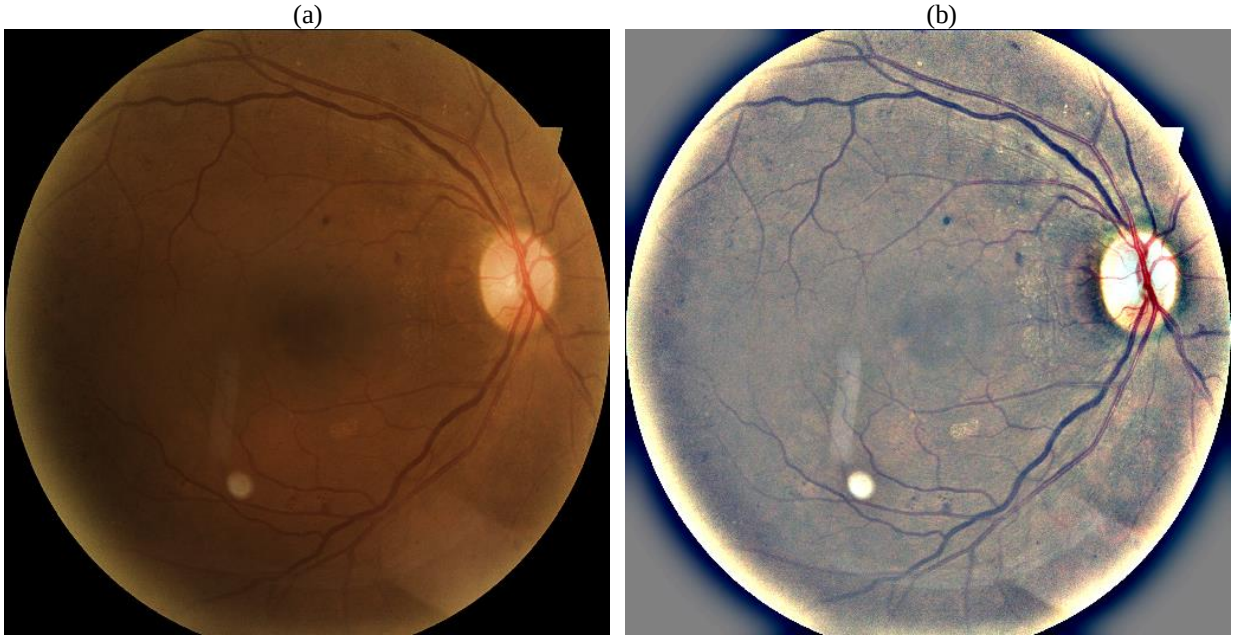
4 https://www.kaggle.com/c/diabetic-retinopathy-detection/data
5 https://medicine.uiowa.edu/eye/abramoff

(I_{label}^{rsd}). To achieve that, each label image (Figure 9) is binarized by a hard threshold – any pixel with a label larger than zero is considered positive. Next, the union of the binarized individual labels is taken as the final pixel-level label:

$$I_{label} = \left(I_{label}^{hem} > 0\right) \cup \left(I_{label}^{rsd} > 0\right). \quad (4)$$

Figure 10 - Pre-processing illustration

(a) (b)



(a) Initial retinal image of the DIARETDB1 dataset and (b) corresponding pre-processed retinal image.
Source: author's own

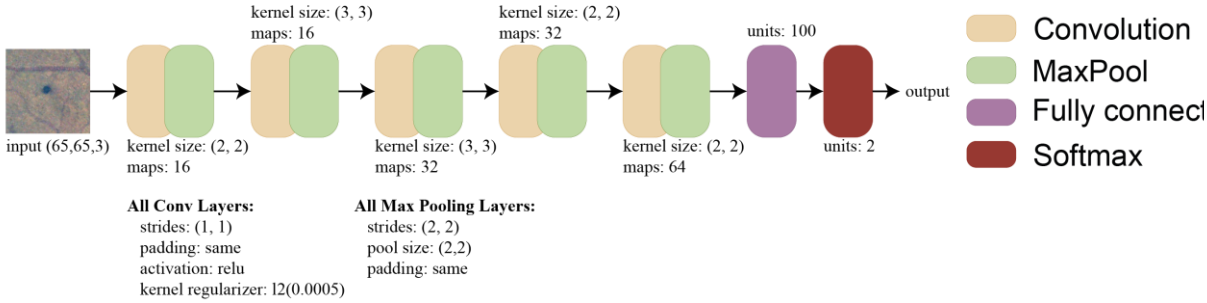
As the database is labeled at the pixel-level, we need a protocol to combine the outputs of the pixels in the patch, thereby producing a single output for each patch. Thus, a patch is considered as a lesion if any pixel in a radius of r pixels around its center is labeled as a lesion. In this work, the values of $p = 65$ and $r = 16$ were used, considering images that are 512 pixels wide.

3.2.3.2 Selecting the patches for fitting the final model

The DIARETDB1 dataset is used as the training set, and its images are divided into training and test sets, as proposed in the database itself. Instead of training and validating the model using all available patches (which would result in more than 23 million samples, considering that the preprocessed images are 512×512 and that there are 262,144 patches per image), a sample selection strategy was used.

We operate in two stages, using two different networks. First, for each image of the training set, all lesion pixels and the same number of random non-lesion pixels are selected to fit a five-layer CNN (Figure 11) called the selection model, for a small number of epochs. This model quickly converges, because the dataset with pre-selected patches is balanced in terms of lesion and non-lesion pixels.

Figure 11 - Selection model architecture



Selection model description. The input is a retinal image patch, and the output is the probability that the input patch contains a red lesion.
Source: author's own

However, this model performs poorly. Indeed, it tends to misclassify some non-lesion patches, as it has not been trained with enough patches with anatomic structures, which may be confused with lesions. We, therefore, use this model to select more challenging non-lesion patches (ones that are misclassified by the selection model) for use in the final model, for performing the last step of transfer learning retraining.

More precisely, we classify all patches of the training set with this selection model. Then, we select all lesion patches and 50,000 non-lesion patches. However, instead of choosing them randomly, they are chosen as follows:

- All non-lesion patches of a retinal image are sorted according to the prediction error of the selection model.
- 50,000 patches with greater prediction errors are used in the final training process.
- The prediction error of each patch is used as a weight in a back-propagation algorithm (sample weight), to give more importance to patches that have been challenging the selection model (where the error was greater) (CIREŞAN et al., 2013).

Using this approach, the final model will be more robust, as it is trained with samples that confused the selection model.

3.2.3.3 Transfer learning

For the final model, we used the Visual Geometry Group (VGG)16 model (SIMONYAN; ZISSERMAN, 2014) because of its high generalization capacity, and because we already have ImageNet pre-trained weights available. The model is initialized with the weights trained on the ImageNet dataset, and is tuned using the patches selected by the selection model as described earlier.

The entire training process is patch-based, as the model is trained using a pair (x_i, y_i) where x_i is an input patch of a retinal image, and y_i is the output label of this patch (as defined in Section 3.2.3.1).

The classification of a pixel is slightly different and is described in the next section.

3.2.4 Calculation of the Lesion Probability Map

It is well known that data augmentation improves the performance of CNNs. We therefore average the model's output of five rotations of the patch, as proposed by Ciresan et al. (CIREŞAN et al., 2013). This process is illustrated in Figure 12.

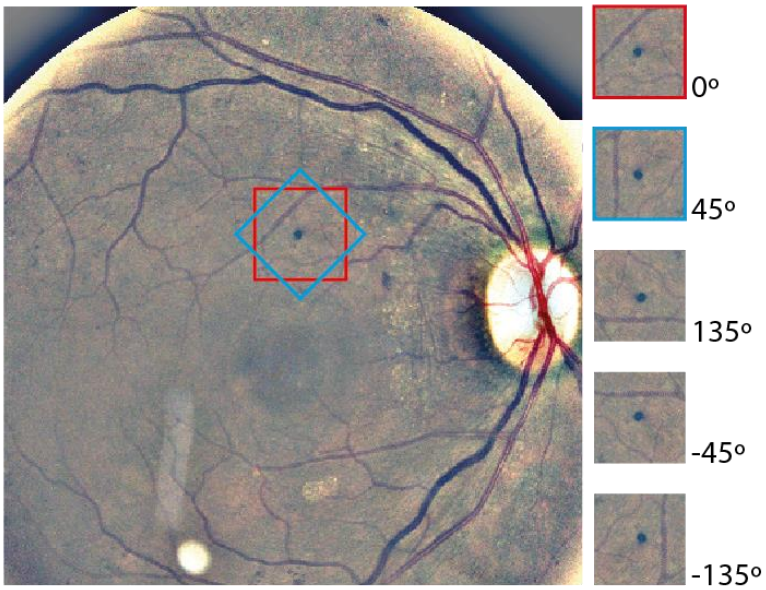
The problem with this approach is that the segmentation of each 512×512 image would require the model to predict $1{,}310{,}720$ patches, which, using a GPU GeForce GTX 1050Ti, would take 20 minutes. A total of 16 days would be required to segment the entire Messidor database.

Considering that the goal is to detect the presence of lesions rather than to provide a precise segmentation, we propose to use a reduced number of patches per image, by considering only part of the image's pixels, as illustrated in Figure 13.

Instead of considering all pixels of the image, we select pixels spaced by S pixels from each other, horizontally and vertically. In this way, the number of patches that must be analyzed in an image is reduced by S^2. Subsequently, the resulting segmented image is resized back to the original dimensions, using extrapolation.

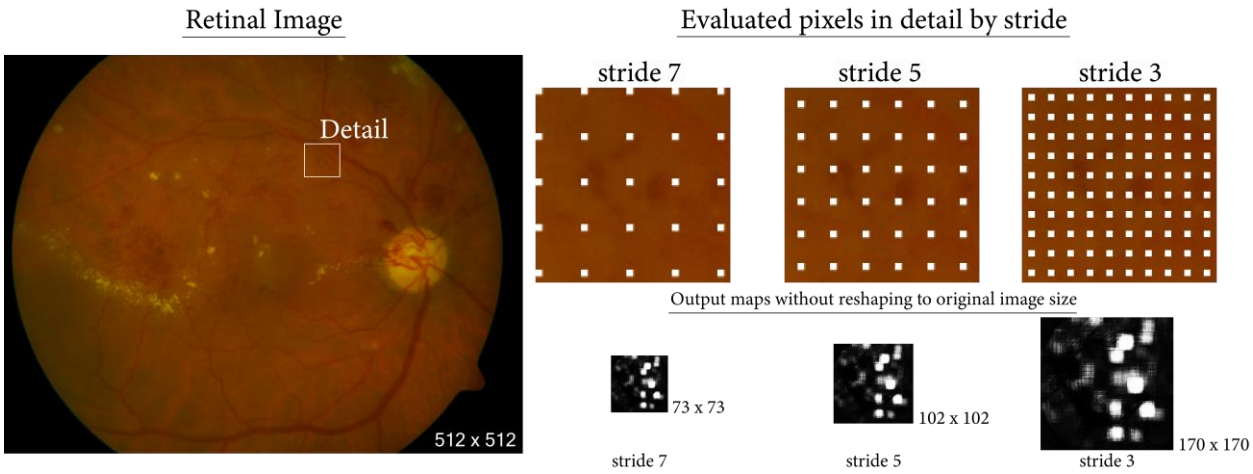
A question that may arise is that using strides could result in missing some lesions. Given that i) a patch is considered as lesion if any pixel in a radius of 16 pixels around its center is labeled as a lesion and, ii) we propose the usage of a 5 pixels stride, even if there is a lesion with 1 pixel of area, which is very unlikely, it would be considered in several patches given the used protocol.

Figure 12 - Patch rotations for lesion prediction



Model predicts the lesions by averaging the output for several rotations of the patch. To generate the rotated patches, the original image is rotated around the center pixel.
Source: author's own

Figure 13 - Use of strides to speed up the segmentation process



Each white square in the images represents an evaluated pixel. The output map is further re-sampled to the input retinal image.
Source: author's own

When the size of the stride increases, the number of patches to be predicted decreases. A 512×512 image that initially required 1,310,720 predictions to be segmented, however

using stride 5, for example, would only require 52,428 predictions, taking only 48 seconds to be processed by the same hardware mentioned before.

The probability map is an image with the same dimensions as the input image with pixel intensity values varying from 0 to 1, representing the probability of a particular pixel being a lesion.

In summary, to obtain the lesion probability map from a retinal image using the trained model, (i) pixels are selected using strides, (ii) the model evaluates patches centered on the chosen pixels, (iii) the output map is resized to the original retinal image size through linear extrapolation.

3.2.5 Inferring the Level of Diabetic Retinopathy (DR) from a Lesion Probability Map

To classify a retinal image (I_{ret}) as indicating DR or not indicating DR, we must infer a single prediction value from its probability map.

Several approaches could be considered for this objective. For example, it would be possible to extract several features from the probability map generated by the model, such as the number of lesions, maximum value, minimum value, the average area of the lesions, and others and to build a new classifier to generate a class (or probability) from those features.

As we want the system to remain generic and straightforward, we propose to represent the image by the maximum value of the probability map, as used in (SEOUD et al., 2016):

$$P(lesion \vee I_{ret}) = max\big(S(I_{ret})\big). \quad (5)$$

Here, $S(I_{ret})$ is the red lesion probability map described above and $P(lesion \vee I_{ret})$ is the probability that a lesion is present in I_{ret}.

Indeed, this quantity:

- tends to increase with the severity of DR; and
- goes from 0 to 1 (or can be monotonically transformed to this range).

3.2.6 Performance Indicators for Diabetic Retinopathy (DR) Detection

The classifier's performances were described using the area under the receiver's operating characteristic curve (AUC), sensitivity (Se), and specificity (Sp).

Contrary to many authors, who present the average level of the performance indicator calculated over the entire dataset, we calculate the confidence intervals (CIs) using the following bootstrap method (DELONG; DELONG; CLARKE-PEARSON, 1988):

1. Given the data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, where x_i represents a retinal image, and y_i the corresponding class (0 or 1):
2. N_{boot} bootstrap empirical samples (b^*) are created from the original data.
3. The classifier (C) performance indicator v_C (AUC and mean squared error (MSE)) is calculated for each b_i^* bootstrap empirical sample.
4. The average $\overline{v_C^*}$ is calculated for each b^*, where v_C^* is the performance indicator for the bootstrap sample b^*.
5. As the bootstrap works by approximating the variation, for each bootstrap sample, the variation $\delta_C^* = \overline{v_C^*} - \overline{v_C}$ is calculated, where $\overline{v_C}$ is calculated in the original sample.
6. For a 95% confidence level, the percentiles 2.5 ($p_{.025}(\quad)$) and 97.5 ($p_{.975}(\quad)$) are taken from all δ_C^*.
7. Finally, the CI for the mean of the performance indicator v_C using the classifier C is calculated by $[\overline{v_C} - p_{.975}(\overline{\delta_C^*}), \overline{v_C} - p_{.025}(\overline{\delta_C^*})]$.

This technique can also be used to conduct paired tests to compare C_1 and C_2 ,as follows.

1. N_{boot} bootstrap empirical samples (b^*) are created from the original data.
2. The classifier performance indicator v_{C_i} is calculated for each b_i^* bootstrap empirical sample, for each classifier C_1 and C_2.
3. As there is no guarantee that the distributions are normal, a one-sided Wilcoxon rank test is conducted with the sets v_{C_1} and v_{C_2} to compare the classifiers.

As a result, the hypothesis test is accepted or rejected, and the $p - value$ is returned. We used $N_{boot} = 2000$.

3.3 Experiments

The model was trained on the DIARETDB1 database, and was tested for DR detection on the datasets described in Section 3.2.1. These datasets contain DR labels with different grades, and are also frequently used in the literature for automatic DR detection, thereby allowing comparison with other works. Keras framework has been used in all experiments.

Two experiments are proposed to assess the model's capacity for detecting DR. As the first grades of DR correspond to images with a very mild level of DR that would not require any ophthalmic intervention, researchers have defined two different classification problems.

1. DR screening: in this experiment, the objective is to detect any level of DR, i.e., to discriminate between normal images and any level of DR; this is challenging, as images with mild DR have very few signs of DR.
2. DR need for referral: the goal in this problem is to separate images with a high level of DR, namely to cluster normal and mild DR versus all others.

We use a stride of 5 pixels in all experiments.

3.3.1 Experimental setup

3.3.1.1 Training on standard diabetic retinopathy database, calibration level 1 (DIARETDB1)

Following the method described in Section 3.2, the model was trained on the DIARETDB1 training set, which is composed of 28 retinal images with pixel-level annotations.

The choice of the patch size was made empirically, by observing the lesions of the training set. The patch should be large enough to contain the largest lesions (such as hemorrhages), but not too large; otherwise, the small lesions would occupy a very small area of the patch. Given that 80% of the hemorrhages have a size smaller than 52 pixels and 80% of the red small dots have a size smaller than 22 pixels, we used patches of a size $p = 65$ pixels, and a radius $r = 16$ pixels.

The selection model is a deep convolution network (LECUN et al., 1989). The details of the different layers are given in Figure 11 on page 65. The model was trained for 50 epochs with 25,920 lesion patches, and with the same number of random non-lesion patches. Both sets of patches were augmented by random rotations between 0 and 360 degrees, and with horizontal and vertical flips. This model is used to select the 50,000 most challenging non-lesion patches that will be used in the second stage, i.e., the ones leading to higher error.

In the second stage, a model is built based on the VGG16 architecture (SIMONYAN; ZISSERMAN, 2014), where the last layers are replaced by three dense layers of size 512, 256, and 2, respectively. The model's weights were initialized through the training of the network for another task (object detection on ImageNet dataset) and were tuned using the DIARETDB1 training set for 1,000 epochs and using the stochastic gradient descent (SGD) algorithm with batches of 128 patches.

In the experiments, we used an initial learning rate of 0.01, with a momentum of 0.9. We implemented early stopping after 15 epochs without improvement on the MSE loss, together with a learning rate reduction by a factor of 0.5 after ten epochs without improvement on the validation set (20% randomly-selected patches out of the training set).

3.3.1.2 Tests on DIARETDB1

The DIARETDB1 test set was used to test the model's ability to detect red lesions. In Section 3.2.3.1, we saw that labeling the lesion requires the choice of a threshold. We selected a threshold that maximized the f_1 score, which is the harmonic average between recall and precision ($f_1 = 2 \cdot precision \cdot recall/(precision + recall)$). The selected threshold was 0.39.

Next, the bootstrap method was applied to calculate the distribution and CIs for recall, precision, and f_1 score on the entire test set of DIARETDB1.

3.3.1.3 Tests on standard diabetic retinopathy database, calibration level 0 (DIARETDB0)

DIARETDB0 has only binary annotations on the presence of lesions. To use it for DR detection, a simple protocol was used: images containing labels for red small dots or hemorrhages were considered as indicating DR. Therefore, the only experiment tested on this dataset is the DR screening experiment.

3.3.1.4 Tests on Messidor

Messidor has been extensively used in the literature to evaluate DR detection approaches. As already described, the images are graded with DR levels from R0 (no DR) to R3. The experiments are designed in the following way:

1- DR screening: {R0} versus {R1, R2, R3}; and

2- DR need for referral: {R0, R1} versus {R2, R3}.

3.3.1.5 Tests on Indian diabetic retinopathy image dataset (IDRiD)

The IDRiD dataset separates images in groups ranging from 0 (no apparent DR) to 4 (severe DR), according to the International Clinical Diabetic Retinopathy Scale (WU et al., 2013). In the experiments, we divided the dataset into {0} versus {1,2,3,4} in the DR screening experiment, and {0,1} versus {2,3,4} in the DR need for referral experiment. Even if this

database provides training and test sets, we decided to use only the test set for comparison with other works.

3.3.1.6 Tests on DDR

The DDR dataset uses the DR grades from the International Classification of DR (INTERNATIONAL COUNCIL OF OPHTHALMOLOGY (ICO), 2017) (from 0 to 4), and a specific label for poor-quality images. Those poor-quality images were discarded from the analysis. We divided the dataset using the same groups as with the IDRiD dataset and tested our model on the test set only.

3.3.1.7 Tests on Kaggle

The Kaggle dataset uses the same grades as IDRiD, with the difference being that Kaggle contains two images per subject, i.e., one per eye. We made use of this extra information and carried out the two experiments described above using per-subject analysis as in other datasets. For each subject, we obtained two prediction values, and the maximum one was selected for classification. As the Kaggle dataset was proposed in a competition context, it only provides grades for the training set. For this reason, we applied our approach on the training set only.

3.3.1.8 Tests on Messidor-2

The Messidor-2 dataset also has two images per subject, and the same protocol as in Section 3.3.1.7 was employed. Only two grade levels are available for this dataset[6]: referable subjects, and non-referable subjects. For that reason, only the DR need for referral experiment is applied to this dataset.

3.3.1.9 Testing the effects of image quality in diabetic retinopathy detection

In order to assess the effect of image quality in the proposed DR detection method, we set up an experiment which included in the front end our quality assessment method. We propose a simple experiment: remove the bad quality images from the dataset and compare the results with all images.

The experiment is composed by the following steps:

1. The DR classes distributions of the entire dataset is observed;

---

6 https://medicine.uiowa.edu/eye/abramoff

2. All images from the dataset are classified by our quality assessment method and poor-quality images are removed;
3. The DR detection metrics are calculated using the same bootstrap method described in section 3.2.6 with the difference that the bootstrap samples are taken following the same class distributions found in item 1 for the sake of comparison. Table 7 indicates the number of images versus DR level classified by our method as of poor quality.

Table 7 - Distribution of poor-quality images per dataset

Dataset	DR level	Number of images	Number of poor-quality images	% of poor-quality images in that DR level
Messidor	0	546	34	6.23%
	1	254	7	2.76%
	2	247	19	7.69%
	3	153	34	22.22%
Kaggle	0	20684	2024	9.79%
	1	1927	110	5.71%
	2	4242	702	16.55%
	3	727	141	19.39%
	4	556	277	49.82%

Source: author's own

The objective of this experiment is to test the hypothesis that discarding poor-quality images improves the DR detection performance.

3.3.2 Experimental Results

In the state of the art, most of the works that apply deep learning for DR detection use hundreds of thousands of images to train the model, meaning a huge burden for the experts to label the images accordingly. As a difference from those works, our model was trained using only 28 images from the DIARETDB1 dataset, with pixel-level annotations.

The results for these datasets are shown in Table 8 for the DR screening experiment, and in Table 9 for the DR need for referral experiment. The results are presented in terms of AUC and Se at a fixed Sp of 50%, as performed by other authors (ORLANDO et al., 2018; SÁNCHEZ et al., 2011; SEOUD et al., 2016). The metrics distributions can be seen in Figure 14. From Table 8, we observe that our model provides a Se superior to 80% in all databases.

Given that the international guidelines recommend a Se ranging from 60% to 80% for DR screening (CHAKRABARTI; HARPER; KEEFFE, 2012; ROYAL COLLEGE OF OPTHALMOLOGY, 2019), our results suggest that our approach can be used in a real scenario for DR screening, with the ability to significantly reduce the burden on experts in triage programs.

Table 8 - Results for the DR screening experiment on several datasets

DR screening

Dataset	**AUC (95% CI)**	**Se (95% CI) ($Sp = 50\%$)**
Messidor	.912 (.897–.928)	.940 (.921–.959)
Kaggle*	.764 (.756–.773)	.911 (.800–.823)
IDRiD	.818 (.742–.898)	.841 (.753–.948)
DDR	.848 (.836–.861)	.891 (.875–.908)
DIARETDB0	.786 (.713–.875)	.821 (.743–.969)

* Per subject evaluation

Source: author's own

Table 9 - Results for the DR need for referral experiments on several datasets

DR need for referral

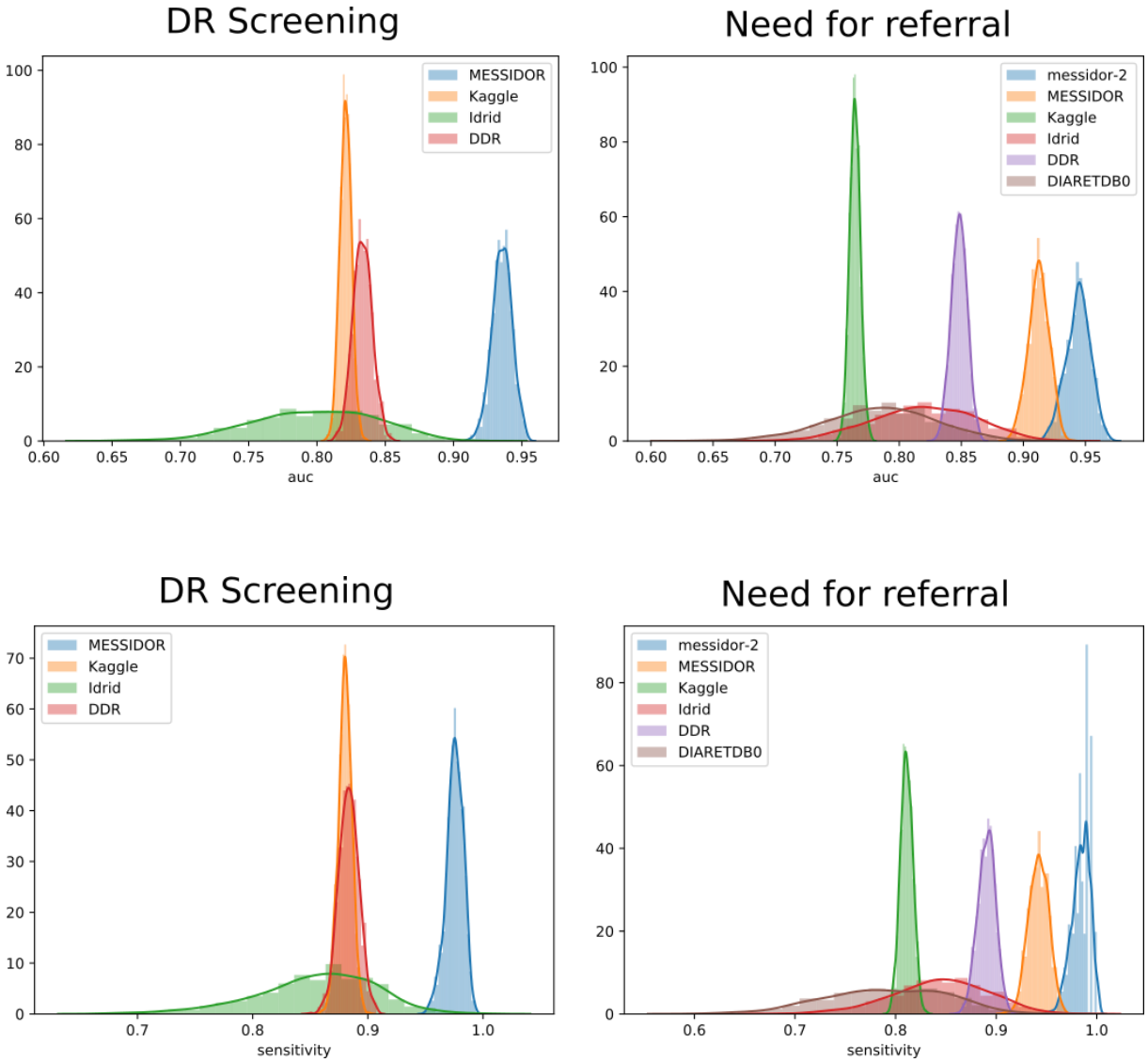
Dataset	**AUC (95% CI)**	**Se (95% CI) ($Sp = 50\%$)**
Messidor	.936 (.921–.950)	.976 (.964–.992)
Kaggle*	.821 (.812–.829)	.881 (.869–.891)
IDRiD	.796 (.715–.892)	.859 (.776–.982)
DDR	.833 (.819–.846)	.885 (.869–.903)
Messidor-2*	.944 (.927–.965)	.984 (.968–1.00)

* Per subject evaluation

Source: author's own

From Figure 14, we observe rather small CIs for all databases, except IDRiD and DIARETDB0. This is owing to their small number of test images, as described in Table 6 on page 63.

Figure 14 - Distributions of area under the receiver's operating characteristic curve (AUC) and sensitivity for both experiments on several datasets



Source: author's own

3.3.3 Comparison with Other Works

To compare our approach with other works, two datasets were employed: Messidor, and Messidor-2. We essentially considered works that tested their results on Messidor or Messidor-2 datasets while training their models on a different dataset (cross-dataset validation), as it

corresponds to our protocol. We also considered the performance of 2 experts. The results are shown in Table 10.

For both experiments on the Messidor dataset, our method resulted in AUC and Se values comparable or superior to those in the state-of-the-art, and even to specialists, indicating that our patch-based CNN approach can be successfully employed both to detect DR and to localize its signs.

We infer that this level of success is owing to several factors. First, the patch selection approach tends to reduce the number of false positives of the classifier, because it enhances the importance of challenging samples.

Second, when compared to Orlando's method (ORLANDO et al., 2018), we note the superiority of learning characteristics adapted to the classification task (rather than using hand-crafted ones), which is also the conclusion of other works regarding medical images.

In this study, we used a single database (DIARETDB1) for localizing the lesion. A question may arise whether using another database would improve the generalization ability, by increasing the diversity of lesion appearance.

However, a careful analysis of the few datasets which present lesion annotations led us not to pursue this direction. Indeed, the Retinopathy Online Challenge (ROC) dataset (NIEMEIJER et al., 2010) labels are not entirely trustworthy, because they result from the union of four specialists' annotations (and not in a grade, as done in DIABRETDB1). From the results of specialists (SÁNCHEZ et al., 2011), it is reasonable to affirm that because they make mistakes such as confusing artifacts with lesions, for example, and using the union of several specialists' annotations tends to increase the number of false-positive labels, leading a model to learn incorrect features.

Another possible public dataset is e-ophta (DECENCIÈRE et al., 2013), but we did not use it because it does not provide hemorrhage labels, which are important markers of DR.

Finally, it seems that the choice of using only the DIARETDB1 dataset for learning was sufficient to ensure a good generalization ability with fast network convergence.

The appropriate statistical method to make a proper comparison between classifiers is a paired test (STAPOR, 2018). Understandably, making a paired test is not always possible, because the results from other approaches must be available. We could do this only with Orlando's et al. model (ORLANDO et al., 2018), as they kindly provided their results for all images of the Messidor database.

Table 10 - Comparison of DR screening and need for referral using the Messidor dataset

Method	DR screening		DR need for referral	
	AUC (95% CI)	Se (95% CI)	AUC (95% CI)	Se (95% CI)
Experts				
Expert A (SÁNCHEZ et al., 2011)	.922 (.902–.936)	.945	.940	.982
Expert B (SÁNCHEZ et al., 2011)	.865 (.789–.925)	.912	.920	.976
Classical image processing				
Sánchez, et al. (SÁNCHEZ et al., 2011)	.876 (.856–.895)	.922	.91	.944
Rocha, et al.[a] (ROCHA et al., 2012)	-	-	.862	-
Giancardo, et al. (GIANCARDO et al., 2013)	.854	-	-	-
Zhang, et al. (ZHANG et al., 2014)	-	-	.930	-
Seoud, et al. (SEOUD et al., 2016)	.899	.939	.916	.962
Deep learning				
Orlando, et al. (ORLANDO et al., 2018)	.893 (.875–.912)	.916 (.894-.943)	.935 (.920–.950)	.974 (.964-1.00)
This work	**.912 (.897–.928)**	**.940 (.921-.959)**	**.936 (.921–.950)**	**.976 (.964-.992)**

[a] {R0} versus {R3}.
[b] Does not use dataset cross-validation

Source: author's own

As the distributions of the result metrics are not normal, a Wilcoxon rank-sum test was applied to the results of both classifiers on the bootstrap sample, and the results can be seen in Figure 14.

Concerning the DR screening experiment, which is the most challenging one, the Wilcoxon rank test (WILCOXON, 1945) for AUC indicated that our approach resulted in an AUC similar to that of Orlando et al. (ORLANDO et al., 2018) ($p > 0.05$) and a Se value greater than Orlando et al. ($p \ll 0.01$). This is probably because of the patch selection procedure that we used for the final training.

Table 11 - Comparison of referable DR detection using the Messidor-2 dataset. Our approach is the only one that employs a patch-based CNN

Method	**AUC (95% CI)**	**Se (95% CI)**	**Sp (95% CI)**	**Number of training images**
Abràmoff, et al. (ABRÀMOFF et al., 2016)	.980 (.968–.992)	.968 (.933–.988)	.870 (.842–.894)	1,250,000
Gulshan, et al. (GULSHAN et al., 2016a)	.990 (.986–.995)	.961 (.924–.983)	.939 (.924–.953)	128,175
Gargeya, et al. (GARGEYA; LENG, 2017)	.940	.930	.87	75,137
This work	.944 (.925–.966)	.900 (.860–.961)	.87 (.863–.871)	28

Source: author's own

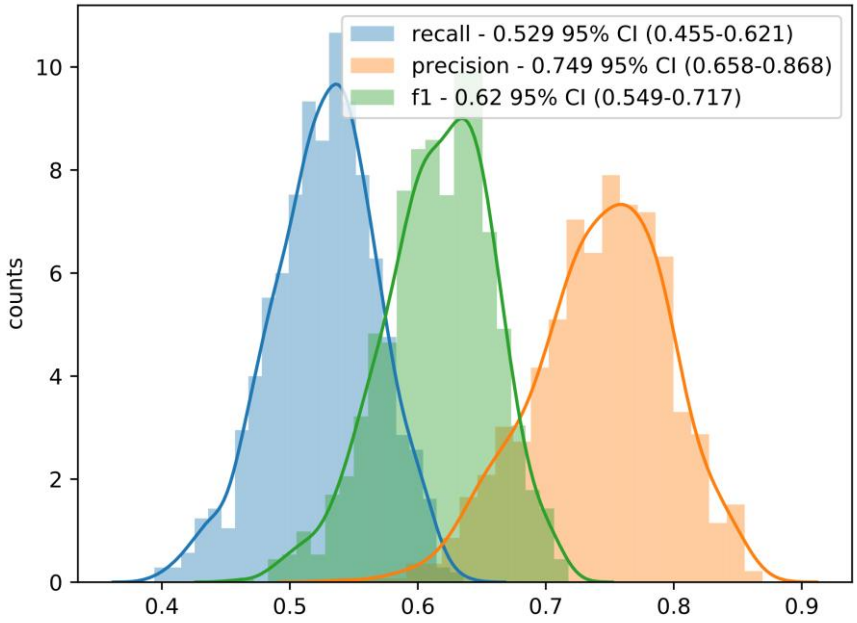
In the Messidor-2 dataset, our approach resulted in an AUC value of 0.944 (95% CI: $0.925 - 0.966$) and an Se value of 0.900 (95% CI: $0.860 - 0.961$) for a fixed Sp of 87%, as shown in Table 11. This fixed Sp point was chosen to compare our work with that of Gargeya et al. (GARGEYA; LENG, 2017). Our results are comparable to Gargeya et al. (GARGEYA; LENG, 2017), with a lighter training phase. Abramoff et al. (ABRÀMOFF et al., 2016) trained with many more images (1.25 million images against 28 images in our training set), but their results in terms of Se and Sp are comparable to ours. In contrast, Gulshan et al. (GULSHAN et al., 2016a) employed both a larger training set (128,175 images) and a different labeling protocol, which in turn produced better results in the Messidor-2 dataset.

Indeed, the work of Gulshan et al. explained through a performance curve that when taking only the annotations of two ophthalmologists, as we did, their performance drops by approximately 30% in terms of Se, which is smaller than ours.

From Table 6 on page 63, we can observe that one of the advantages of the patch-based approach is that it allows for training of the CNN with only a few images with pixel-level annotations. Therefore, for DR classification, we obtain results similar to those of other works that require a massive number of graded images.

Apart from DR detection, our approach also performs lesion localization. We, therefore, designed another experiment on the DIARETDB1 test set to evaluate the performance of our lesion localization. The results are presented in Figure 15. We observe that the recall is relatively low. However, because of the redundancy of lesions (there is generally more than one lesion per image), the DR screening results are good, as shown in Figure 15.

Figure 15 - Performance metric of the lesion detection in the DIARETDB1 test set



Source: author's own

Figure 16 shows a qualitative result of lesion localization on an image from the Messidor database, i.e., a false-positive (left) detection, and a true-positive (right) detection, as obtained by our model.

We can observe that the localization of red lesions produced by the system can be useful for specialists to focus their attention on a limited region of the image, thereby reducing the burden of evaluating the retinal image globally. Apart from using the complete model for DR detection, the system can also be used for potential lesion detection. In that case, a lower threshold can be used to increase the model's recall of image regions that might contain signs of DR.

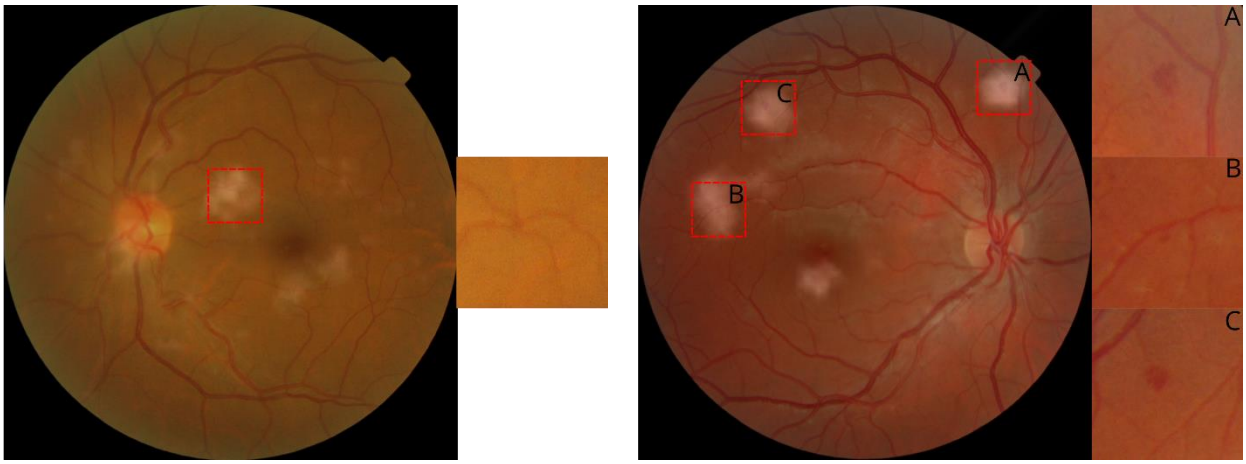
It is important to emphasize that manual image assessment is a difficult task even for experts, as it depends on small lesion localization in retinal images.

3.3.3.1 The effect of image quality in diabetic retinopathy detection

The two most commonly used public retinal images datasets with DR labels have been chosen to be used in this experiment: Messidor and Kaggle.

As expected, using only good-quality images improves the performance of the DR detection method as shown in Figure 17. In order to confirm through a statistical method that using only good-quality images improves the DR detection, a two-sample Z test has been applied resulting in $p_{value} \ll 0.001$ for both Messidor and Kaggle datasets.

Figure 16 - Qualitative results of the red lesion localization



(a) (b)

Both images are composed of the original retinal image with the output of our approach superimposed and the localized lesions highlighted. The regions delimited by red squares are shown in detail at the right side of the retinal image. Figure (a) shows a false-positive example, which is an intersection between blood vessels and (b) three true-positive examples.
Source: author's own

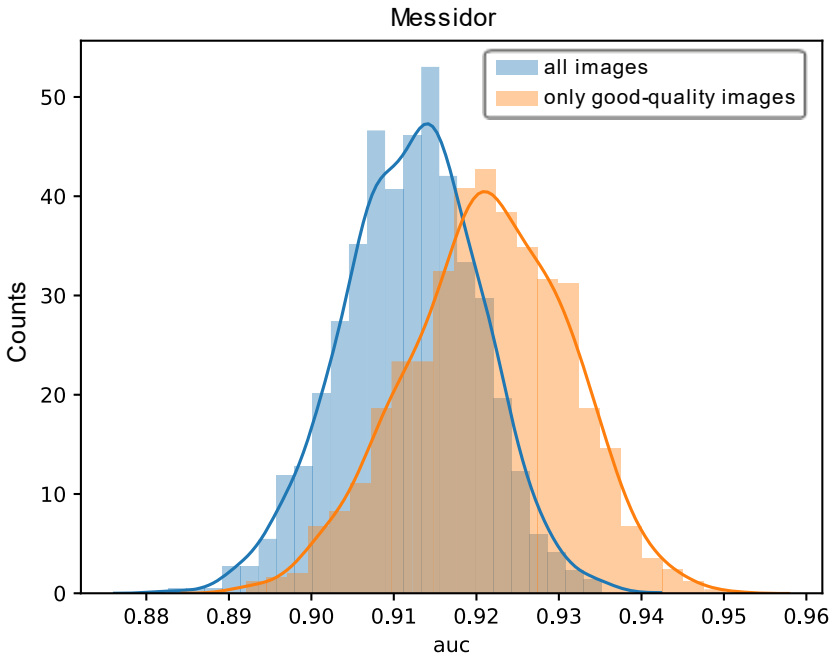
Although significantly different distributions have been found, the DR detection method might have some robustness to poor-quality images because it works on patches – a poor-quality image might have some regions with good quality – and also because the model has been trained without removing the poor-quality images from the dataset.

This way, the methods based on classical image processing might be more benefited from the removal of poor-quality images.

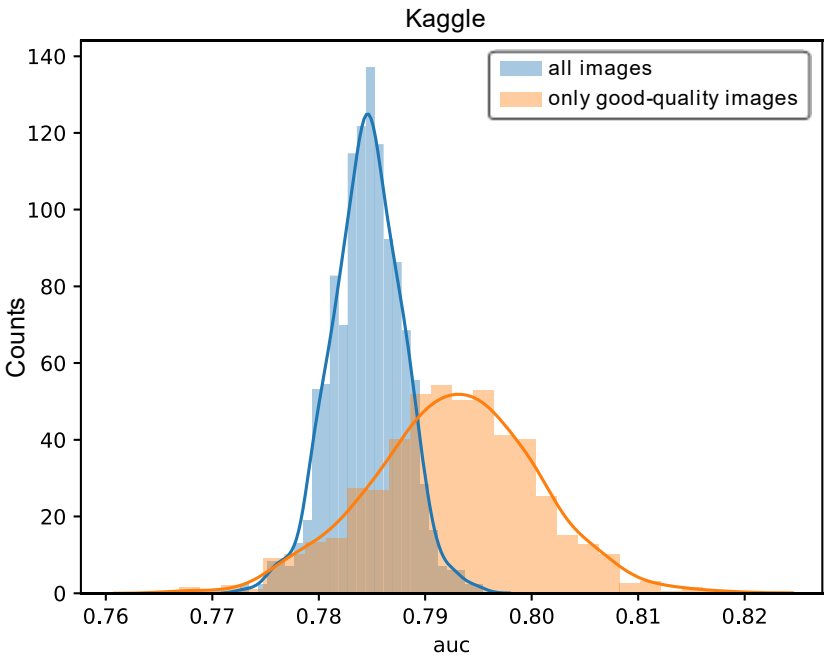
3.4 Conclusion

The present study is devoted to the early detection of DR in retinal images using a CNN deep network approach. In contrast to many authors who use CNN globally on the pixels of the image to produce a label, our model deals with patches and can localize potential regions of lesions, providing a powerful tool for a specialist in retinal red lesion detection, and leading to further DR detection.

Figure 17 - Area under the curve (auc) distribution for diabetic retinopathy detection concerning the removal of poor-quality images



(a)



(b)

Area under the curve (auc) distribution comparison for (a) Messidor and (b) Kaggle datasets using all images and only good quality images.
Source: author's own

Our main concern was to reduce the complexity of the model, while improving its performance. To this end, we designed an efficient procedure for selecting training patches, so that challenging examples would be given special attention during the training process. Moreover, as our aim was a crude localization of the regions rather than their precise segmentation, we under-sampled the images, leading to a drastic reduction in the processing time of generalization, which is important in a real application scenario.

The classification decision is based on a probability map indicating the level of DR of each pixel. We used a simple procedure to deduce a label function for the complete image, with the advantage of not requesting any retraining step. More precisely, our approach separates the phase of lesion localization (learned by a CNN) and DR labeling as a direct outcome of the outputs of the CNN. In this way, our model can be used immediately on any database, without any further adaptation.

Extensive experiments on several databases have shown that our approach outperforms other models tested under the same experimental conditions, namely, in a cross-validation context where the test and training databases are different.

Moreover, in contrast to the state-of-the-art, training our model requires only a small number of images, which is of high value considering the burden of labeling large quantities of images.

We also showed that removing the poor-quality images might improve the performance of the proposed DR detection pipeline.

Future research could investigate building separate models for each type of lesion such as MAs, hemorrhages, and even bright lesions such as exudates. This would allow other databases to be used, such as e-ophta or DDR, and would allow us to make use of their large number of images. Another direction would be to consider other types of data (such as clinical data) to complement the data image to improve the recognition performance, rather than using costly specialist labeling.

4 FINAL CONCLUSIONS AND FUTURE WORKS

In this thesis two issues involving retinal image processing have been studied. The first one is the quality assessment of retinal images and the second is the localization of red lesions which leads to DR detection.

We used transfer learning on a deep neural network to assess quality of retinal images. This allowed a reduction of the images needed to train the model and reduced the time spent on this phase. Two limitations of this part of the work are the small number of public datasets with quality labels and the small number of poor-quality images in those datasets. However, cross-dataset validation showed that the model developed is generic enough to be applied on images from different devices, which is very important since mobile retinal cameras are becoming more popular.

Next, we proposed a patch-base lesion localization model in order to detect images with DR. Different from many other approaches that work on the image level, we decided to work at the pixel level providing a powerful tool for the specialist of retinal red lesion detection, leading to further DR detection. As our main goal was to provide a crude localization of the lesions rather than precisely segment them, it allowed us to greatly reduce the computational cost by under-sampling the evaluated patches while preserving the both the lesion localization, which is important for the specialist and the DR detection capability, important for usage in triage of DR patients. To ensure that our model was not dataset dependent we employed dataset cross-validation using several datasets and results similar to the state-of-the-art approaches was found. One of the advantages of our approach is the reduced number of labelled images needed in the training phase. While other approaches have been trained with hundreds of thousands of even more than a million images, our model has been trained using only 28 retinal images.

Therefore, we have confirmed our hypothesis through experiments that it is possible to propose a pipeline based on quality assessment and red lesion localization to achieve automatic DR detection with performance similar to experts considering that a rough segmentation is sufficient to produce a discriminant marker of a lesion have been confirmed.

4.1 Perspectives

Several works showed that DR detection models have performance similar to medical experts. An ensemble of models can be used to annotate retinal images datasets that don't possess proper DR labels such as UK-Biobank and ELSA-Brasil.

Both datasets mentioned above have retinal images of the same patients taken in different times which allows to evaluate the disease development though the analyses of the retinal lesions and other clinical data as well, allowing to investigate the factors that influence in the development of DR.

A recent work proposed the use of Generative Adversarial Networks (GAN) models to synthetize retinal images (COSTA et al., 2018). Different from other GAN models, their approach uses the blood vessels map as input in order to create a new retinal image from it. One of the problems is that for each blood vessels map a single image is created since there is no random variable as input as done in other GAN models.

Adding stochasticity, i.e., allowing the model to create different images from the same blood vessels map, would greatly enhance the model`s usage to create big synthetic retinal images datasets.

Another improvement would be to insert not only the blood vessels map as input but also lesions maps (of several types of lesions) allowing the model to create retinas with and without lesions. With that, models with much greater capacity would be able to be trained without manual annotations from the experts.

4.2 Published works

The publications resulting from this research, in chronological order, are listed below:

1. ZAGO, Gabriel Tozatto e colab. **Retinal image quality assessment using deep learning. Computers in Biology and Medicine**, v. 103, n. September, p. 64–70, Dez 2018.
2. ZAGO, Gabriel Tozatto e colab. **Diabetic retinopathy detection using red lesion localization and convolutional neural networks**. Computers in Biology and Medicine, v. 116, p. 103537, 1 Jan 2020.

5 REFERENCES

ABDEL-HAMID, L. et al. Retinal image quality assessment based on image clarity and content. **Journal of Biomedical Optics**, v. 21, n. 9, p. 096007, 16 set. 2016.

ABDEL HAMID, L. S. S. et al. No-reference wavelet based retinal image quality assessment. **Computational Vision and Medical Image Processing V - Proceedings of 5th Eccomas Thematic Conference on Computational Vision and Medical Image Processing, VipIMAGE 2015**, p. 123–130, 2016.

ABRÀMOFF, M. D. M. D. et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. **Investigative Opthalmology & Visual Science**, v. 57, n. 13, p. 5200, 4 out. 2016.

AHMAD FADZIL, M. H. et al. Analysis of retinal fundus images for grading of diabetic retinopathy severity. **Medical and Biological Engineering and Computing**, v. 49, n. 6, p. 693–700, 2011.

ALGHADYAN, A. A. **Diabetic retinopathy - An updateSaudi Journal of Ophthalmology**, 2011.

ANTAL, B.; HAJDU, A. Improving microaneurysm detection using an optimally selected subset of candidate extractors and preprocessing methods. **Pattern Recognition**, v. 45, n. 1, p. 264–270, 1 jan. 2012.

AQUINO, E. M. L. L. et al. Brazilian Longitudinal Study of Adult health (ELSA-Brasil): Objectives and design. **American Journal of Epidemiology**, v. 175, n. 4, p. 315–324, 15 fev. 2012.

AZIZPOUR, H. et al. **From generic to specific deep representations for visual recognition**. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. **Anais**...IEEE, jun. 2015

BALASUBRAMANIAN, S.; PRADHAN, S.; CHANDRASEKARAN, V. **Red lesions detection in digital fundus images**. 2008 15th IEEE International Conference on Image Processing. **Anais**...IEEE, 2008

BARCELÓ, A. et al. The cost of diabetes in Latin America and the Caribbean. **Bulletin of the World Health Organization**, v. 81, n. 1, p. 19–27, 2003.

BARTLING, H.; WANGER, P.; MARTIN, L. Automated quality evaluation of digital fundus photographs. **Acta Ophthalmologica**, v. 87, n. 6, p. 643–647, 2009.

CHAKRABARTI, R.; HARPER, C. A.; KEEFFE, J. E. Diabetic retinopathy

management guidelines. **Expert Review of Ophthalmology**, v. 7, n. 5, p. 417–439, 9 out. 2012.

CHUDZIK, P. et al. Microaneurysm detection using fully convolutional neural networks. **Computer Methods and Programs in Biomedicine**, v. 158, p. 185–192, maio 2018.

CHUI, T. Y. P. et al. The mechanisms of vision loss associated with a cotton wool spot. **Vision Research**, v. 49, n. 23, p. 2826–2834, 2009.

CIREŞAN, D. C. et al. Mitosis detection in breast cancer histology images with deep neural networks. **Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention**, v. 16, n. Pt 2, p. 411–8, 2013.

COSTA, P. et al. End-to-End Adversarial Retinal Image Synthesis. **IEEE Transactions on Medical Imaging**, v. 37, n. 3, p. 781–791, 2018.

DAVIS, H. et al. **Vision-based, real-time retinal image quality assessment**. 2009 22nd IEEE International Symposium on Computer-Based Medical Systems. **Anais**...IEEE, ago. 2009

DECENCIÈRE, E. et al. TeleOphta: Machine learning and image processing methods for teleophthalmology. **IRBM**, v. 34, n. 2, p. 196–203, 1 abr. 2013.

DECENCIÈRE, E. et al. Feedback on a publicly distributed image database: The Messidor database. **Image Analysis and Stereology**, v. 33, n. 3, p. 231–234, 2014.

DELONG, E. R.; DELONG, D. M.; CLARKE-PEARSON, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. **Biometrics**, v. 44, n. 3, p. 837, set. 1988.

ETDRS. Grading DR from Stereoscopic Color fundus Photographs - An extension of the Modified Airlie Housa Classification - ETDRS 10. **Ophthalmology**, v. 98, p. 786–806, 1991.

EWING, D. J.; CLARKE, B. F. Diagnosis and management of diabetic autonomic neuropathy. **British Medical Journal**, v. 285, p. 916–918, 1982.

FAUST, O. et al. **Linear and non-linear analysis of cardiac health in diabetic subjectsBiomedical Signal Processing and Control**, 2012a.

FAUST, O. et al. Algorithms for the Automated Detection of Diabetic Retinopathy Using Digital Fundus Images: A Review. **Journal of Medical Systems**, v. 36, n. 1, p. 145–157, fev. 2012b.

GARGEYA, R.; LENG, T. Automated Identification of Diabetic Retinopathy Using Deep Learning. **Ophthalmology**, v. 124, n. 7, p. 962–969, jul. 2017.

GIANCARDO, L. et al. **Elliptical local vessel density: A fast and robust quality metric for retinal images**. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. **Anais**...IEEE, ago. 2008

GIANCARDO, L. et al. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. **Medical Image Analysis**, v. 16, n. 1, p. 216–226, 1 jan. 2012.

GIANCARDO, L. et al. **Validation of microaneurysm-based diabetic retinopathy screening across retina fundus datasets**. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. **Anais**...IEEE, jun. 2013

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [s.l.] MIT Press, 2016.

GOODFELLOW, I. J. et al. Generative Adversarial Networks. **JAMA Internal Medicine**, v. 177, n. 3, p. 1–9, 10 jun. 2014.

GRUNDY, S. M. et al. Diabetes and cardiovascular disease: a statement for healthcare professionals from the American Heart Association. **Circulation**, v. 100, n. 10, p. 1134–1146, 1999.

GULSHAN, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. **JAMA - Journal of the American Medical Association**, v. 316, n. 22, p. 2402–2410, 2016a.

GULSHAN, V. et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. **JAMA**, v. 316, n. 22, p. 2402, 13 dez. 2016b.

GUTHRIE, D. W.; GUTHRIE, R. A. **The Diabetes Sourcebook**. 5th. ed. Los Angeles: Lowell House, 2003.

HARMAN-BOEHM, I. et al. The eyes in diabetes and diabetes through the eyes. **Diabetes Research and Clinical Practice**, v. 78, n. 3, p. S51–S58, dez. 2007.

HARNEY, F. Diabetic retinopathy. **Medicine**, v. 34, n. 3, p. 95–98, mar. 2006.

HUNTER, A. et al. **An automated retinal image quality grading algorithm.Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference**, 2011.

INTERNATIONAL COUNCIL OF OPHTHALMOLOGY (ICO). **International**

Council of Ophthalmology Guidelines for Glaucoma Eye Care, 2017.

JIA DENG et al. **ImageNet: A large-scale hierarchical image database**. 2009 IEEE Conference on Computer Vision and Pattern Recognition. **Anais**...2009

KANDE, G. B. et al. **Detection of red lesions in digital fundus images**. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. **Anais**...IEEE, jun. 2009

KANDE, G. B.; SAVITHRI, T. S.; SUBBAIAH, P. V. Automatic Detection of Microaneurysms and Hemorrhages in Digital Fundus Images. **Journal of Digital Imaging**, v. 23, n. 4, p. 430–437, 17 ago. 2010.

KAUPPI, T. et al. DIARETDB0 : Evaluation Database and Methodology for Diabetic Retinopathy Algorithms. **Machine Vision and Pattern Recognition Research Group, Lappeenranta University of Technology, Finland.**, p. 1–17, 2006.

KAUPPI, T. et al. **the DIARETDB1 diabetic retinopathy database and evaluation protocol**. Procedings of the British Machine Vision Conference 2007. **Anais**...2007

LALONDE, M.; GAGNON, L.; BOUCHER, M. Automatic visual quality assessment in optical fundus images. **Vision Interface (VI2001)**, p. 259–264, 2001.

LARSEN, M. et al. Automated detection of fundus photographic red lesions in diabetic retinopathy. **Investigative Ophthalmology and Visual Science**, v. 44, n. 2, p. 761–766, 1 fev. 2003.

LECUN, Y. et al. Backpropagation Applied to Handwritten Zip Code Recognition. **Neural Computation**, 1989.

LEE, S. C.; WANG, Y. Automatic retinal image quality assessment and enhancement. **Proceedings of SPIE Image Processing.**, v. 3661, n. February, p. 1581–1590, 1999.

LI, T. et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. **Information Sciences**, v. 501, p. 511–522, out. 2019.

LITJENS, G. et al. A survey on deep learning in medical image analysis. **Medical Image Analysis**, v. 42, p. 60–88, 1 dez. 2017.

MACGILLIVRAY, T. J. et al. Suitability of UK Biobank retinal images for automatic analysis of morphometric properties of the vasculature. **PLoS ONE**, v. 10, n. 5, p. e0127914, 22 maio 2015.

MAHAPATRA, D. et al. Retinal image quality classification using saliency maps and CNNs. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [s.l: s.n.]. v. 10019 LNCSp.

172–179.

MOOKIAH, M. R. K. et al. Computer-aided diagnosis of diabetic retinopathy: A review. **Computers in Biology and Medicine**, v. 43, n. 12, p. 2136–2155, dez. 2013.

NAIR, V.; HINTON, G. E. **Rectified Linear Units Improve Restricted Boltzmann Machines**. Proceedings of the 27th International Conference on Machine Learning (ICML). **Anais**...2010

NIEMEIJER, M. et al. Automatic detection of red lesions in digital color fundus photographs. **IEEE Transactions on Medical Imaging**, v. 24, n. 5, p. 584–592, maio 2005.

NIEMEIJER, M. et al. Retinopathy Online Challenge: Automatic Detection of Microaneurysms in Digital Color Fundus Photographs. **IEEE Transactions on Medical Imaging**, v. 29, n. 1, p. 185–195, jan. 2010.

NIEMEIJER, M.; ABRÀMOFF, M. D.; VAN GINNEKEN, B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. **Medical Image Analysis**, v. 10, n. 6, p. 888–898, 2006.

NUGROHO, H. A. et al. Contrast measurement for no-reference retinal image quality assessment. **Proceedings - 2014 6th International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology Through University-Industry Collaboration, ICITEE 2014**, p. 8–11, out. 2015.

OPHTHALMIC PHOTOGRAPHERS' SOCIETY. **Ophthalmic Photographers' Society**.

ORGANIZATION, W. H. **Diabetes. Fact sheet N°312. August 2011World Health Organization**, 2011.

ORLANDO, J. I. et al. An ensemble deep learning based approach for red lesion detection in fundus images. **Computer Methods and Programs in Biomedicine**, v. 153, p. 115–127, 1 jan. 2018.

PAULUS, J. et al. Automated quality assessment of retinal fundus photos. **International Journal of Computer Assisted Radiology and Surgery**, v. 5, n. 6, p. 557–564, nov. 2010.

PENATTI, O. A. B.; NOGUEIRA, K.; DOS SANTOS, J. A. **Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?** IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. **Anais**...IEEE, jun. 2015

PIRES DIAS, J. M.; OLIVEIRA, C. M.; DA SILVA CRUZ, L. A. Retinal image quality assessment using generic image quality indicators. **Information Fusion**, v. 19, n. 1, p. 73–90, set. 2014.

PORWAL, P. et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. **Data**, v. 3, n. 3, p. 25, jul. 2018a.

PORWAL, P. et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. **Data**, v. 3, n. 3, p. 25, 10 jul. 2018b.

QUELLEC, G. et al. Deep image mining for diabetic retinopathy screening. **Medical Image Analysis**, v. 39, p. 178–193, 2017.

QUELLEC, G. G. et al. Optimal filter framework for automated, instantaneous detection of lesions in retinal images. **IEEE Transactions on Medical Imaging**, v. 30, n. 2, p. 523–533, fev. 2011.

RAZAVIAN, A. S. et al. **CNN features off-the-shelf: An astounding baseline for recognition**. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. **Anais**...IEEE, jun. 2014

ROCHA, A. et al. Points of Interest and Visual Dictionaries for Automatic Retinal Lesion Detection. **IEEE Transactions on Biomedical Engineering**, v. 59, n. 8, p. 2244–2253, ago. 2012.

ROYAL COLLEGE OF OPTHALMOLOGY. **Diabetic eye screening: commission and provide**.

SAHA, S. K. et al. Automated Quality Assessment of Colour Fundus Images for Diabetic Retinopathy Screening in Telemedicine. **Journal of Digital Imaging**, p. 1–10, 27 abr. 2018.

SÁNCHEZ, C. I. et al. Evaluation of a Computer-Aided Diagnosis System for Diabetic Retinopathy Screening on Public Data. **Investigative Opthalmology & Visual Science**, v. 52, n. 7, p. 4866, 30 jun. 2011.

SAÚDE, M. DA S. (BRASIL). S. DE P. DE; ESTRATÉGICAS., D. DE A. P. **Plano de Reorganização da Atenção à Hipertensão arterial e ao Diabetes mellitus Manual de Hipertensão arterial e Diabetes mellitus**. [s.l: s.n.].

SCANLON, P. H. **Diabetic retinopathyMedicine**, 2010.

SEOUD, L. et al. Red Lesion Detection Using Dynamic Shape Features for Diabetic Retinopathy Screening. **IEEE Transactions on Medical Imaging**, v. 35, n. 4, p. 1116–1126, abr. 2016.

SEVIK, U. et al. Identification of suitable fundus images using automated quality assessment methods. **Journal of Biomedical Optics**, v. 19, n. 4, p. 046006, 9 abr. 2014.

SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. p. 1–14, 4 set. 2014.

STAPOR, K. Evaluating and comparing classifiers: Review, some recommendations and limitations. In: KURZYNSKI, M.; WOZNIAK, M.; BURDUK, R. (Eds.). . **Advances in Intelligent Systems and Computing**. Cham: Springer International Publishing, 2018. v. 578p. 12–21.

SZEGEDY, C. et al. **Rethinking the Inception Architecture for Computer Vision**. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. **Anais**...2016

TAJBAKHSH, N. et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? **IEEE Transactions on Medical Imaging**, v. 35, n. 5, p. 1299–1312, maio 2016.

TING, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. **JAMA - Journal of the American Medical Association**, v. 318, n. 22, p. 2211–2223, 2017.

UK BIOBANK. **http://www.ukbiobankeyeconsortium.org.uk**.

USHER, D. B.; HIMAGA, M.; DUMSKYJ, M. J. Automated assessment of digital fundus image quality using detected vessel area. **Proceedings of Medical Image Understanding and Analysis**, p. 81–84, 2003.

VALLABHA, D. et al. Automated detection and classification of vascular abnormalities in diabetic retinopathy. **Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.**, v. 2, p. 1625–1629, 2004.

VAN GRINSVEN, M. J. J. P. J. P. et al. Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images. **IEEE Transactions on Medical Imaging**, v. 35, n. 5, p. 1273–1284, maio 2016.

WELIKALA, R. A. A. et al. Automated retinal image quality assessment on the UK Biobank dataset for epidemiological studies. **Computers in Biology and Medicine**, v. 71, p. 67–76, 1 abr. 2016.

WILCOXON, F. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**,

1945.

WILD, S. et al. Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030. **Diabetes Care**, v. 27, n. 5, p. 1047–1053, 2004.

WILKINSON, C. P. et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. **Ophthalmology**, 2003.

WILLIAMS, R. et al. Epidemiology of diabetic retinopathy and macular oedema: A systematic review. **Eye**, v. 18, n. 10, p. 963–983, 2004.

WU, L. et al. Classification of diabetic retinopathy and diabetic macular edema. **World Journal of Diabetes**, v. 4, n. 6, p. 290, 15 dez. 2013.

YAO, Z. et al. **Generic features for fundus image quality evaluation**. 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services, Healthcom 2016. **Anais**...IEEE, set. 2016

YEN, G. G.; LEONG, W.-F. W. F. A sorting system for hierarchical grading of diabetic fundus images: A preliminary study. **IEEE Transactions on Information Technology in Biomedicine**, v. 12, n. 1, p. 118–130, jan. 2008.

YU, H. et al. **Automated image quality evaluation of retinal fundus photographs in diabetic retinopathy screening**. 2012 IEEE Southwest Symposium on Image Analysis and Interpretation. **Anais**...IEEE, abr. 2012

ZAGO, G. T. et al. Retinal image quality assessment using deep learning. **Computers in Biology and Medicine**, v. 103, n. September, p. 64–70, dez. 2018.

ZAGO, G. T. et al. Diabetic retinopathy detection using red lesion localization and convolutional neural networks. **Computers in Biology and Medicine**, v. 116, p. 103537, 1 jan. 2020.

ZHANG, B. et al. Hierarchical detection of red lesions in retinal images by multiscale correlation filtering. **Proceedings of SPIE**, v. 7260, p. 72601L-72601L–12, 26 fev. 2009.

ZHANG, B. et al. Detection of microaneurysms using multi-scale correlation coefficients. **Pattern Recognition**, v. 43, n. 6, p. 2237–2248, jun. 2010.

ZHANG, X. et al. Exudate detection in color retinal images for mass screening of diabetic retinopathy. **Medical Image Analysis**, v. 18, n. 7, p. 1026–1043, 1 out. 2014.