Alberto Iglesias Burgos

ASSESSING DIABETIC RETINOPATHY BY MEANS OF LESIONS DETECTION IN EYE-FUNDUS IMAGES

FINAL DEGREE PROJECT

Directed by Aïda Valls Mateu

Degree in Computer Engineering

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2023

Acknowledgements

I would like to express my gratitude to Aïda Valls for her guidance during the completion of this final degree project. Her insights and support were instrumental in the development of this project.

I would also like to thank Eugeni and Salem for their valuable collaboration and guidance at various stages of this work. Their expertise and contributions significantly enriched the project.

I also wish to acknowledge the invaluable contribution of Dr. Pedro Romero and Dr. Marc Baget, distinguished ophthalmologists, who provided valuable medical insights, generated relevant images, and offered valuable guidance in result validation.

This work is part of the research project ADRIANA, funded by the Carlos III Health Institute and the European Union (PI21/00064).

Resum.

La retinopatia diabètica és una de les principals causes de discapacitat visual evitable, que afecta principalment la població en edat laboral a nivell mundial. Els recents avanços han destacat la necessitat de mètodes més eficients i econòmics per a facilitar la identificació i el diagnòstic precoç de malalties a la retina. Tenint en compte la importància dels programes de detecció de la retinopatia diabètica i les dificultats associades amb l'obtenció de diagnòstics primerencs fiables, aquest projecte explora la viabilitat d'emprar eines de diagnòstic assistit per ordinador.

Aquest projecte investiga la possibilitat de diagnosticar la gravetat de la retinopatia diabètica mitjançant l'anàlisi d'imatges que contenen diverses lesions, incloent-hi microaneurismes i hemorràgies. Múltiples *datasets* públics que contenen imatges de fons d'ull de pacients seran sotmesos a una anàlisi automàtica utilitzant models de *deep learning*.

El flux de treball integra l'anàlisi d'imatges realitzat tant per experts humans com per models de *deep learning*, utilitzant específicament el model LezioSeg, per identificar i quantificar diferents lesions. Posteriorment, els comptatges de lesions es processen utilitzant diversos mètodes per determinar el grau de retinopatia diabètica en cada pacient. Es compararan els resultats obtinguts per oftalmòlegs humans amb els generats pel sistema, avaluant d'aquesta manera l'eficàcia i fiabilitat del model.

Pel fet que els resultats obtinguts amb dades públiques no van complir amb els estàndards desitjats, es va elaborar minuciosament un nou *dataset* en col·laboració amb l'Hospital Universitari Sant Joan de Reus, que es va emprar per a tornar a entrenar el model LezioSeg. Aquest procés va tenir com a objectiu millorar el model, per abordar així els reptes inicials.

Resumen.

La retinopatía diabética es una de las principales causas de discapacidad visual evitable, que afecta principalmente a la población en edad laboral a nivel mundial. Los recientes avances han destacado la necesidad de métodos más eficientes y económicos para facilitar la identificación y el diagnóstico precoz de enfermedades retinianas. Teniendo en cuenta la importancia de los programas de detección de la retinopatía diabética y las dificultades asociadas con la obtención de diagnósticos tempranos fiables, este proyecto explora la viabilidad de emplear herramientas de diagnóstico asistido por ordenador.

Este proyecto investiga la posibilidad de diagnosticar la gravedad de la retinopatía diabética mediante el análisis de imágenes que contienen diversas lesiones, incluyendo microaneurismas y hemorragias. Múltiples *datasets* públicos que contienen imágenes de fondo de ojo de pacientes serán sometidos a un análisis automático utilizando modelos de *deep learning*.

El flujo de trabajo integra el análisis de imágenes realizado tanto por expertos humanos como por modelos de *deep learning*, utilizando específicamente el modelo LezioSeg, para identificar y cuantificar diferentes lesiones. Posteriormente, los conteos de lesiones se procesan utilizando diversos métodos para determinar el grado de retinopatía diabética en cada paciente. Se

compararán los resultados obtenidos por oftalmólogos humanos con los generados por el sistema, evaluando de esta manera la eficacia y fiabilidad del modelo.

Debido a que los resultados obtenidos con datos públicos no cumplieron con los estándares deseados, se elaboró minuciosamente un nuevo *dataset* en colaboración con el Hospital Universitario Sant Joan de Reus, que se empleó para volver a entrenar el modelo LezioSeg. Este proceso tuvo como objetivo mejorar el modelo, para abordar así los desafíos iniciales.

Abstract.

Diabetic retinopathy is a leading cause of preventable vision impairment, primarily affecting the global working-age population. Recent advancements in research have underscored the need for more efficient and cost-effective methods to facilitate the early identification and diagnosis of retinal diseases. Recognizing the significance of diabetic retinopathy screening programs and the challenges associated with achieving reliable early diagnoses, this project explores the feasibility of employing computer-aided diagnostic tools.

This project explores the feasibility of diagnosing the severity of diabetic retinopathy by analysing images containing diverse lesions, including microaneurysms and haemorrhages. Multiple publicly datasets containing retinal fundus images from patients, will be subject to automatic analysis utilizing deep learning models.

The workflow integrates image analysis conducted by both human experts and deep learning models, specifically utilizing the LezioSeg model, to identify and quantify various types of lesions. Subsequently, the lesion counts are further processed using diverse methodologies to determine the severity of diabetic retinopathy in each patient. The study will then compare the results obtained by human ophthalmologists and the artificial intelligence system, to evaluate the efficacy and reliability of the deep learning approach.

However, initial results obtained with public data from this approach did not meet the desired standards. Consequently, a new dataset was meticulously curated in collaboration with the Hospital Universitario Sant Joan de Reus, which was used to retrain the LezioSeg model. This process aimed to enhance the model's capabilities and address the challenges encountered in the initial analysis.

Contents

List of Tables

List of Figures

1 Introduction

This Bachelor thesis is framed into a research project conducted at the ITAKA research group of the Department of Computer Science and Mathematics at Universitat Rovira i Virgili (URV) together with the Ophthalmology research group from URV and IISPV (Institut d'Investigació Sanitària Pere Virgili).

Since 2010 the groups are working in improving the screening of Diabetic Retinopathy and provide greater health coverage to risky people. This project has been financially supported by the *Investigación en Salud, Instituto de Salud Carlos III* and *Fondos Feder*. Through several funded projects, they have been developing a computerized system to aid clinical diagnosis for the screening of diabetic retinopathy. The system should assist family doctors in the detection of the personalized risk of developing diabetic retinopathy. There are two study lines, one focused on automatic image analysis and a second specialized on the personalized detection of DR risk by means of different clinical values stored in the Electronic Health Record. Some more information can be found in the webpage: https://deim.urv.cat/~itaka/retiprogram/

This Bachelor thesis will contribute to the line of image analysis to improve the diagnosis of Diabetic Retinopathy. The rest of the chapter presents the background concepts of this disease as well as the basics of the Artificial Intelligence methods used in this work.

1.1 Diabetic Retinopathy

Diabetic retinopathy (DR) is a frequent and distinctive microvascular complication that arises from diabetes, and it continues to be the primary, avoidable cause of blindness among individuals of working age. With the global prevalence of diabetes mellitus (DM) on the rise, DR remains a significant contributor to vision impairment in numerous developed nations. While diabetes can impact the eye in various ways, such as increasing the risk of cataracts, diabetic retinopathy stands out as the most prevalent and severe ocular complication. In 2010, of the 246 million people with diabetes, about a third have signs of diabetic retinopathy, and a third of these might have vision-threatening retinopathy, defined as severe retinopathy or macular oedema [1]. According to World Health Organization, 422 million people suffer from DR, with the number of patients expected to reach epidemic levels worldwide in the next few decades, estimated to reach 642 million by 2040. Most of patients with diabetes mellitus type-1 and about 12% of patients with diabetes type-2 will develop DR at some moment [2]. This work will focus only on the case of Type-2 diabetes mellitus .

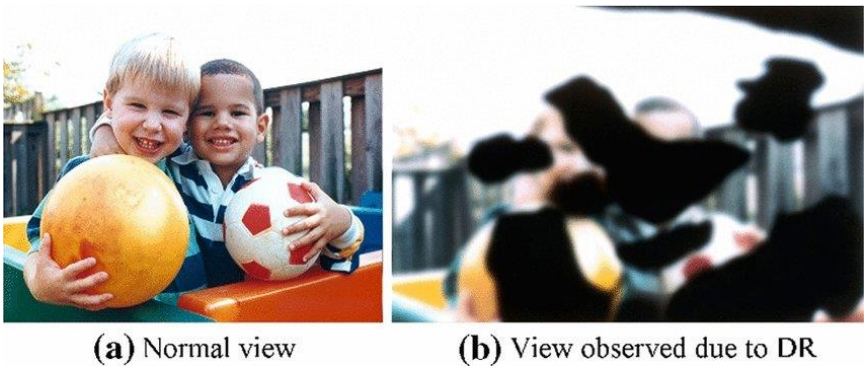


Figure 1. Example of normal vision (left) versus vision with diabetic retinopathy (right). [3]

Diabetic retinopathy will remain a common complication of DM and a leading cause of preventable blindness in the adult working population [4]. It is important to note that not every patient who develops diabetic retinopathy will experience severe vision loss. Vision loss occurs only in advanced stages typified by diabetic macular oedema (DME) and/or proliferative diabetic retinopathy (PDR).

Diabetes is thought to affect many, perhaps all, of the more than 30 cell types in the retina, but because the inner retinal vasculature is so readily visualized, the classification and grading of retinopathy has been based on the severity of vascular lesions. Commonly, the study of DR is based on the anatomical features of the retina and on the number of photographically detectable microvascular lesions. It must be noted that being anatomically-based, this type of study is not a true quantitative measure, and may not reflect important functional deficits. The fact that retinopathy is 'geographically disperse' across an individual retina, and between the two eyes of an individual person, bring added challenges in obtaining quantitative data for analysis in clinical research. As technologies improve, there is an opportunity for improved assessments that more accurately define retinopathy and its effects.

In order to prevent the damage of this severe complication to patients' vision, it is very important to diagnose diabetic retinopathy and provide appropriate treatment to minimize further deterioration as early as possible [5].

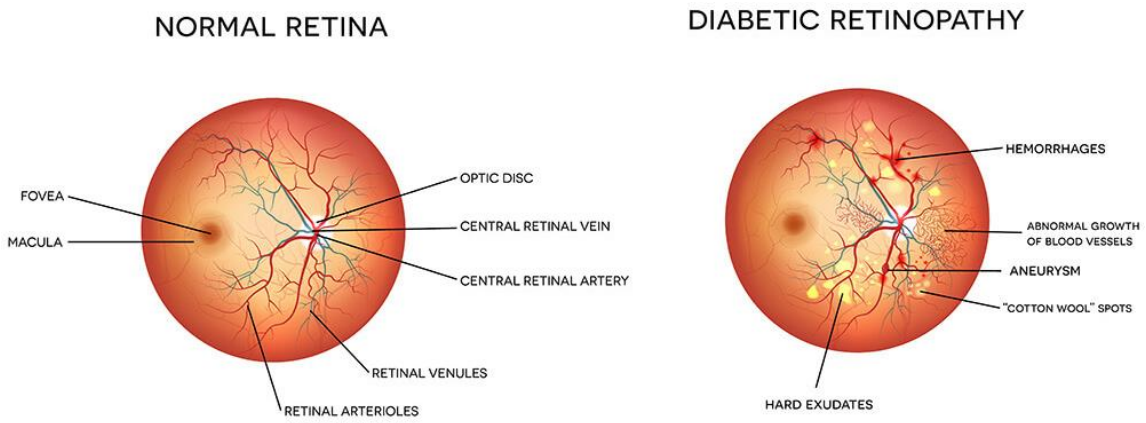


Figure 2. Example of normal retina (left) versus retina with diabetic retinopathy (right). [6]

Diabetic retinopathy may be very broadly classified into two stages based on the level of microvascular degeneration and related ischemic damage: non-proliferative diabetic retinopathy (NPDR) and advanced, proliferative diabetic retinopathy, which was mentioned before. NPDR can be sub-classified into mild, moderate and severe NPDR. The progression of diabetic retinopathy is related to abnormalities of the vasculature including permeability of the blood retina barrier, progressive microvascular damage with vascular endothelial cell and pericyte loss, subsequent occlusion of capillaries and excessive retinal neuronal and glial abnormalities [7].

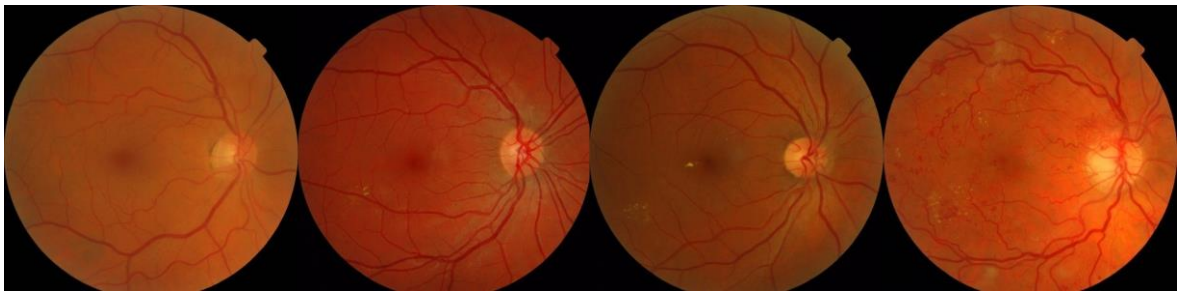


Figure 3. Examples of fundus images, arranged from left to right, depict the absence of retinopathy and the three stages of retinopathy in increasing severity.

Automated DR screening from fundus retinal images, also called retinographies, can be performed by detecting abnormalities such as microaneurysms (MA), haemorrhages (HM), hard and soft exudates (HE and SE) and neovascularization. Based on the detected abnormalities, a patient can be classified as healthy or affected by DR. The method used will be based on counting two of the lesions mentioned above, microaneurysms and haemorrhages [8]. MA and HM are also termed as red lesions. These structures are mainly found in retinal images affected by DR and hypertension. Due to the similar appearance of MA and HM, even for clinicians it is difficult to differentiate between these two structures [9].

1.1.1 Microaneurysms

Retinal microaneurysms are usually the first visible sign of diabetic retinopathy, but also present in other pathologies that affect micro-vessels. Microaneurysms are a small widening of capillary walls. It is not clear whether retinal microaneurysms are caused by damage to blood vessel walls or the onset of neovascularization. However, the result is the appearance of small saccular structures, with approximate dimensions of between 10 and 100μm, that in in colored fundus images appear as round red spots. They are indistinguishable from small bleeding with the same dimensions, because both are small round areas, with a dark red colour [10].

Figure 4. Marked microaneurysms in a fundus image.

Therefore, both microaneurysms and bleeding are smaller than the main venous calibre at the optical disk margin (usually 125μm), considered to be a red dot, and evaluated as microaneurysms. Conversely, any red spots larger than that are considered bleeding, except the features like a round shape, smooth margins and central light reflexes indicate that it might be a microaneurysm [10].

Since microaneurysms are the first signs of DR, its detection is vital. It is also crucial to monitor the development of the disease and classify changes in retinal images [10].

1.1.2 Hemorrhages

Retinal bleeding is deposits of blood in the retina. Bleeding disappears when blood is absorbed again over time. They are caused by rupture of blood vessel walls or microaneurysms, and their increased presence is a clear sign of the use of a damaged retina. They have very different shapes, ranging from round red dots with sharp margins, to spotting bleeding, to bleeding in the form of fire. When blood is reabsorbed, the hemorrhagic margin fades and the distinctive red color turns grayish red before disappearing completely [10].

Figure 5. Marked haemorrhages in a fundus image.

All in all, retinal images are a key factor for ophthalmologists in the diagnosis of DR, being periodic ophthalmoscopy the best approach for eye disease screening. However, the number of ophthalmologists available is a limiting factor in initiating screening. This whole process could be enhanced using automatic analysis of digital images. On the other hand, computer science, especially the field of deep learning could identify objects in many domains, one of which is the identification of objects in images, including images of the retina of the eye.

1.2 Deep learning-based segmentation

Advancements in medical technologies have supported the universal goal of optimizing the efficiency of healthcare systems. Computer vision-based applications are gaining more importance in the field of biomedical imaging, providing decision support information of value, and enhancing the diagnosis. In the specific application field of retinal imaging, different image modalities can be used for the analysis and treatment of DR [11].

Computer-aided diagnosis is an important component of medical informatization. Classification, identification, and segmentation of lesions based on medical imaging are critical for disease follow-up diagnosis and treatment plan formulation. Deep learning techniques, particularly convolutional neural networks (CNN), are quickly becoming the ideal solution for automated medical lesion recognition, thanks to the outstanding advances of Artificial Intelligence (AI) technologies represented by deep learning in natural picture processing [12]. Deep learning-based approaches have proven to be quite effective in single lesion recognition and segmentation. Multiple-lesion recognition is more difficult than single-lesion recognition due to the little variation between lesions or the too wide range of lesions involved.

1.2.1 Deep Learning

To put Deep Learning in context, it is necessary to explain certain related concepts. First of all, there is the more general term AI, which is the field that studies how to develop software systems that exhibit a behavior that one will associate with an intelligent human. For example, the capacity of learning and discovering new knowledge from robust datasets to enable problem-solving. AI employs predictions and automation to optimize and solve complex tasks that humans have historically done.

Machine Learning, a subset of Artificial Intelligence, is a crucial component that enhances the capabilities of AI systems. Unlike traditional programming, where explicit instructions are provided to perform specific tasks, Machine Learning enables systems to learn from data and improve their performance over time [13]. This ability to learn from experience is what sets machine learning apart and empowers AI systems to handle complex tasks effectively.

Deep Learning is a subset of Machine Learning that focuses on using a software model known as neural networks to process vast amounts of data. These neural networks are designed to mimic the structure and functioning of the human brain, allowing them to extract intricate patterns and representations from the data. Deep Learning has gained immense popularity and success in recent years, achieving remarkable results in various domains such as computer vision, natural language processing, and speech recognition. The relationship between these concepts is hierarchical, Deep Learning is a specific technique within the broader domain of Machine Learning, which, in turn, falls under the umbrella of Artificial Intelligence.

Overall, the progression from AI to machine learning and then to Deep Learning reflects the evolution of increasingly sophisticated methods to create intelligent systems that can analyze and interpret data, learn from it, and ultimately perform complex tasks that were previously only achievable by humans. These technologies, in combination, have opened new possibilities and revolutionized numerous industries, making AI a powerful and transformative force in today's world.

1.2.2 Convolutional Neural Network

Convolutional neural network has been making brilliant achievements. It has become one of the most representative neural networks in the field of deep learning. Computer vision based on CNN has enabled people to accomplish milestones, such as face recognitions, autonomous vehicles, self-service supermarkets, and intelligent medical treatments.

The CNN is a kind of feedforward neural network that is able to extract features from data with convolution structures. Different from the traditional feature extraction methods, CNN does not need to extract features manually. The architecture of CNN is inspired by visual perception. A biological neuron corresponds to an artificial neuron; CNN kernels represent different receptors that can respond to various features; activation functions simulate the function that only neural electric signals exceeding a certain threshold can be transmitted to the next neuron. Loss functions and optimizers are something people invented to teach the whole CNN system to learn what we expect. Compared with fully connected (FC) networks in *Figure 6*, CNN possesses many advantages [14].

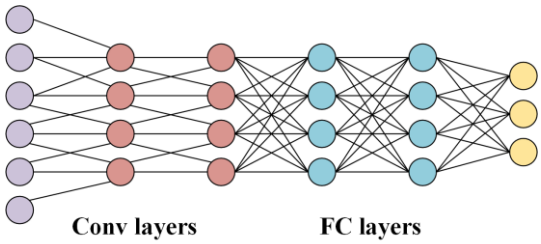


Figure 6. Diagram of CNN layers and FC layers. [14]

For example, this type of network only has local connection. Which means, that each neuron is no longer connected to all neurons of the previous layer, just to a small number of neurons, which is effective in reducing parameters and speed up convergence. These convolutional layers learn local patterns, and these patterns are searched for by means of small 2D kernels by means of small windows or 2D kernels that traverse the input.

Also, can happen weight sharing phenomenon, so a group of connections can share the same weights, which reduces parameters further. As CNNs are primarily used for image classification and segmentation, and it works by finding similar patterns throughout the input. These patterns can be found by sliding a filter with shared weights across the input. The shared weights concept allows the network to learn the same pattern, regardless of its position in the input. CNNs employ multiple filters to find different patterns in the input, which leads to a feature map [15].

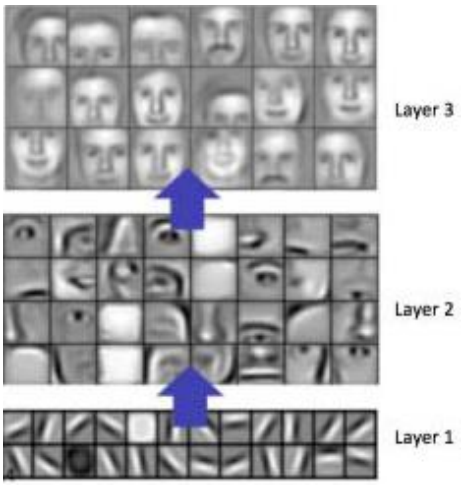


Figure 7. Learned features from a Convolutional Neural Network. [15]

It should also be noted the down-sampling dimension reduction using pooling layers, which consist of a layer that can reduce the amount of data while retaining useful information. It can also reduce the number of parameters by removing trivial features [15].

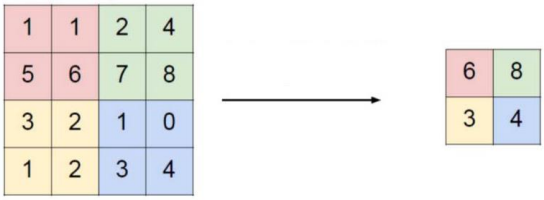


Figure 8. Max-pooling is demonstrated. The max-pooling with 2x2 filter and stride 2 lead to down-sampling of each 2x2 blocks is mapped to 1 block (pixel). [15]

These three appealing characteristics make CNN one of the most representative algorithms in the deep learning field.

1.2.2.1 Convolutional Layer

The convolutional layer is the core building block of a CNN, and it is where most of the computation occurs. It requires a few components, which are input data, a filter, and a feature map. Let's assume that the input will be a colour image, which is made up of a matrix of pixels in 3D. This means that the input will have three dimensions (a height, width, and depth) which correspond to RGB in an image. There is a feature detector, also known as a kernel or a filter, which will move across the receptive fields of the image, checking if the feature is present. This process is known as a convolution [16].

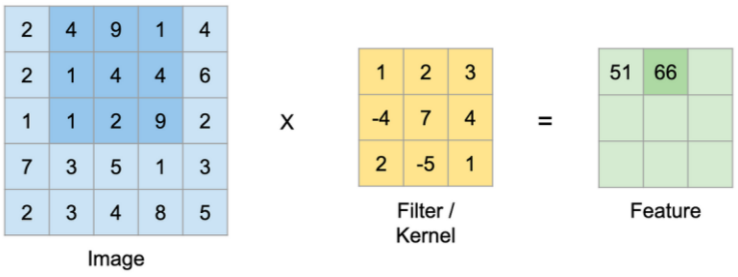


Figure 9. Convolution operation. [17]

The feature detector is a two-dimensional (2-D) array of weights, which represents part of the image. While they can vary in size, the filter size is typically a 3x3 matrix; this also determines the size of the receptive field. The filter is then applied to an area of the image, and a dot product is calculated between the input pixels and the filter. This dot product is then fed into an output array. Afterwards, the filter shifts by a stride, repeating the process until the kernel has swept across the entire image. The final output from the series of dot products from the input and the filter is known as a feature map, activation map, or a convolved feature.

After each convolution operation, a CNN applies an activation function, usually it is used Rectified Linear Unit (ReLU) transformation, to the feature map, introducing nonlinearity to the model.

As it is mentioned earlier, another convolution layer can follow the initial convolution layer. When this happens, the structure of the CNN can become hierarchical as the later layers can see the pixels within the receptive fields of prior layers. As an example, let's assume that we're trying to determine if an image contains a bicycle. You can think of the bicycle as a sum of parts. It is comprised of a frame, handlebars, wheels, pedals, et cetera. Each individual part of the bicycle makes up a lower-level pattern in the neural net, and the combination of its parts represents a higher-level pattern, creating a feature hierarchy within the CNN.

1.2.3 Phases of the Machine Learning Process

In the context of the use of neural networks and the process of machine learning, it is essential to understand the various stages involved in their effective application.

Initially, data preparation and model design are conducted. The data preparation phase involves collecting and preparing the data needed to train and evaluate the model The data includes the observations that we wish to predict or classify, as well as the labels or true values that will be used as a reference to measure the performance of the model. These data are divided into training, validation, and test sets for later use. Model design involves selecting and configuring the type of machine learning model to be used. This

includes the choice of algorithms, the architecture of the neural network in the case of neural networks, and the configuration of key hyperparameters.

Once all the necessary “materials” are ready to use, the training of the model is carried out. Using the training data set, the model adjusts its internal parameters over multiple iterations. As the training process progresses, the model “learns” from this data, identifying patterns and relationships that allow it to make accurate predictions on new, unseen data.

After each training session, the model can undergo a validation phase, where an independent validation dataset is used to assess its performance. This phase helps to find potential problems of over- or under-fitting of the model and allows the hyperparameters to be adjusted if necessary.

Finally, the model is exhaustively evaluated using a test dataset that it has not seen before. This supplies a realistic estimate of its performance in real situations. Evaluation metrics are obtained, which show the accuracy and effectiveness of the model on the specific task.

Once the model has been successfully trained, validated and tested, it is ready for the inference phase. At this stage, the model is used in real applications to make predictions on new data. It can make decisions, classify data or make predictions based on its previous training.

1.2.4 Image Segmentation

Computer vision is a branch of artificial intelligence that deals with enabling machines to interpret and understand visual information from the world, similar to how humans perceive and interpret images. It involves developing algorithms and models that allow computers to process and analyze visual data, such as images and videos. The goal of computer vision is to enable machines to extract meaningful information from visual inputs and make intelligent decisions based on that information.

Currently, a wide variety of computer vision techniques can be found, including, image classification, object detection and segmentation. For this project, the aim is to identify the different lesions that may be present in a retinal fundus image. Being a segmentation problem.

Image segmentation is a fundamental task within computer vision. It refers to the process of dividing an image into multiple segments or regions, where each segment represents a distinct object, region, or area of interest within the image. The objective of segmentation is to group pixels or elements of an image that share common visual characteristics, such as color, texture, or shape. Like many other areas of computer vision, research on segmentation has received a tremendous performance boost with the emergence of deep learning in recent years.

Figure 10. Types of segmentation, from left to right: semantic segmentation, instances segmentation and panoptic segmentation. [18]

By employing Deep Learning techniques for image segmentation, computer vision systems can achieve more accurate and robust results across various applications. Deep learning models can learn to recognize and delineate intricate object boundaries, making them highly valuable tools in computer vision and related fields. In the case of retinal lesion detection, all the above tools in synergy can help, improve, and speed up this process.

2 Hypothesis and Goals

The **aim** of this bachelor thesis is to facilitate the integration of progressive deep learning-based computer vision techniques in the medical field, to assist in identifying ocular lesions and thereby diagnosing diabetic retinopathy.

The **hypothesis** of this project is that, by using a deep learning model it is possible to segment and count the number of different lesions in retinal fundus images, from which it is possible to correctly diagnose the grade of diabetic retinopathy present in the patient.

As indicated in *Chapter 1*, this bachelor thesis has been developed in the frame of a Spanish research project at the ITAKA research group. In 2022, a deep learning system called LezioSeg was developed to generate binary masks of different DR lesions from retinal fundus images [19]. The LezioSeg system is the starting point of this work.

The workflow involves image analysis by both human experts and deep learning models to identify and quantify the distinct types of lesions. The obtained lesion counts are further processed using various methods, to figure out the severity of diabetic retinopathy in each patient. The study will compare the results obtained by human ophthalmologists and the AI system to evaluate the efficacy and reliability of the deep learning approach. Additionally, the project seeks also to assess the viability of this system in a dataset from a local hospital in Catalonia.

In order to achieve this hypothesis, the following goals can be defined.

- Define (with ophthalmologists) a set of rules to determine the degree of DR from the number of Microaneurysms and Haemorrhages found in different regions of the eye fundus image.
- Integrate these rules with the LezioSeg deep learning model into a unique diagnosis system.
- Carry out an extensive testing and comparative analysis of the results obtained by human ophthalmologists and the diagnostic system to evaluate the efficacy and reliability of the proposal.

From these general goals, it is possible to identify some more specific sub-goals:

- To get knowledge of ocular lesions and pathologies and their diagnosis.
- To learn and use image processing and data manipulation libraries to extract and store the necessary information from the available images.
- To learn concepts related to machine learning, neural networks, and image segmentation, including the libraries and source code of the LezioSeg implementation made at ITAKA group.
- To make a flexible implementation of the different rules based on the number of different types of lesions, in different regions of the image from the inference results given by LezioSeg.
- To prepare several datasets to assess the implemented DR diagnosis system.
- To compare and analyse the results of the predictions and diagnoses with the data provided by experts in the field, the latter being considered as ground truth (GT).
- To carry out the retraining of the neural network with a new data set in order to improve the accuracy of the predictions in the future. The new data set will have to be created from scratch.

3 Methodology

3.1 Indian Diabetic Retinopathy Image Dataset database

The Indian Diabetic Retinopathy Image Dataset (IDRiD) database was available as a part of "Diabetic Retinopathy: Segmentation and Grading Challenge" organized in conjunction with IEEE[1] International Symposium on Biomedical Imaging (ISBI-2018). The IDRiD dataset consist of 516 high-resolution fundus images categorized in retinal images with the signs of DR and/or DME and normal retinal images (without signs of DR and/or DME), being jpg format of 4288 × 2848 pixels [20].

The dataset is composed of three subsets, which may contain part of the images mentioned above. The creation of these subsets was to achieve the following tasks: pixel level annotation/segmentation to locate the four types of lesions (microaneurysms, haemorrhages, hard exudates, and soft exudates) and the optic disc; disease grading of diabetic retinopathy, and the third objective was optic disc and fovea centre coordinates. The first two subsets should be emphasised, as they are the ones that will be used.

3.1.1 Pixel Level Annotated Data

This dataset consists of 81 colour fundus images with signs of DR. Precise pixel level annotation as shown in *Figure 11* of abnormalities associated with DR like microaneurysms, soft exudates, hard exudates and haemorrhages is provided as a binary mask for performance evaluation of individual lesion segmentation techniques. It includes colour fundus images and separate binary masks for each lesion type (.tif files). Along with the lesion masks, it also consists of optic disc (OD) mask for all 81 images. The dataset is divided into 54 images as a training set and the rest of 27 as a testing set [20].

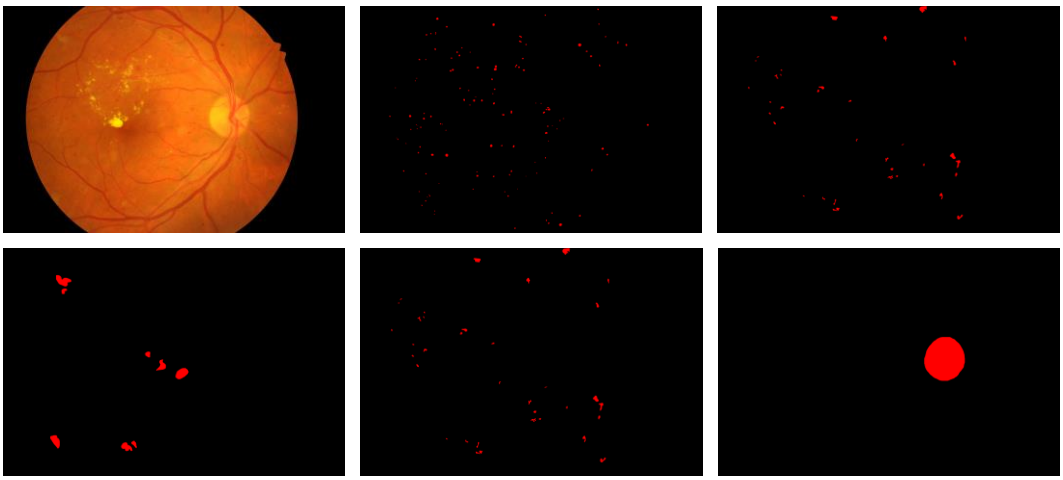


Figure 11. Fundus image and its respective ground truth masks, from left to right and top to bottom: microaneurysms, haemorrhages, soft exudates, hard exudates and optic disc. [20]

3.1.2 Image Level Disease Grading

The medical experts graded the full set of 516 images with a variety of pathological conditions of DR and DME [20]. The expert labels of DR and DME severity level for the dataset were saved on a CSV file with each column description given as follows:

[1] Institute of Electrical and Electronics Engineers

- Image number: name (serial number) of deidentified and renamed patient image.
- DR grade: DR severity level in range 0 (no apparent DR) to 4 (severe DR).
- Risk of DME: DME severity level in range 0 (no DME) to 2 (severe DME).

3.2 Messidor Database

This database was created within the Messidor project to evaluate different lesion segmentation methods for colour eye fundus images, in the framework of diabetic retinopathy screening and diagnosis. The set consists of 1200 fundus images, originally 640 × 640 pixels in size. These images were organised into different directories according to the degree of retinopathy diagnosed by the expert. However, this dataset did not include any associated masks [21].

3.3 Sant Joan de Reus Hospital Dataset

The private dataset of the Sant Joan de Reus Hospital, which was created for this project, consists of a total of 111 images, which have a jpg format and a dimension of 1261 × 939 pixels. In this case, starting from the total set of images, two different sets were created, one having the masks manually labelled by the hospital for the microaneurysms' lesions (101 images/masks) and another for the haemorrhage's lesions (111 images/masks). The fact that the MA dataset is smaller is because those masks with no lesion were removed. Each dataset is divided into training, validation and testing set making up of around 80%, 10% and 10% images and their masks respectively, by maintaining appropriate mixture of disease stratification.

From both datasets a variation was made in which the images and masks were split into four equal parts, creating larger datasets but where the images are smaller in dimension.

3.4 LezioSeg

LezioSeg is a state-of-the-art model to create an accurate lesion segmentation from fundus images based on deep learning approaches. It has been developed in the ITAKA research group, and it is the main output of the PhD thesis of Dr. Mohammed Yousef Salem Ali. This model contains two multi-scale modules to enhance deep-learning segmentation model performance for extracting relevant features from the eye fundus images. The first multi-scale module is used at the bottleneck of the LezioSeg network. The second multi-scale attention (SAT) module is used with the decoder to capture a wider range of relevant features by mixing low and high-resolution data from different decoder layer sources, to enhance the concentration of the small objects that might be lost while the image reconstruction in the decoder block. Another important novelty, it is the integration of a gated skip connections (GSCs) mechanism at the decoder of LezioSeg to help the network focus on retinal lesion features coming from the encoder [19].

As a result, the computational cost is much lower than the backbone-based ones like ResNets, VGGs, and DensNets or those that depend on more than one backbone encoder.

The architecture of LezioSeg is composed of three parts, as shown in *Figure 12*. First, the encoder network (i.e., the backbone) encodes the input image and generates feature maps. Second, there is an Atrous Spatial Pyramid Pooling (ASPP) layer after the encoder network (i.e., the neck) that can capture contextual information at multiple scales for generating better representations of the small lesions of the retinal eye. Third, the decoder network (i.e., the head) contains four blocks, each having a GSCs mechanism to

encourage the model to learn eye lesions-relevant features [22]. Finally, a multi-scale attention mechanism (the one called SAT) is connected with each decoder block as an additional lesion segmentation boost to enhance learning efficiency by combining low and high-resolution data from different sources.

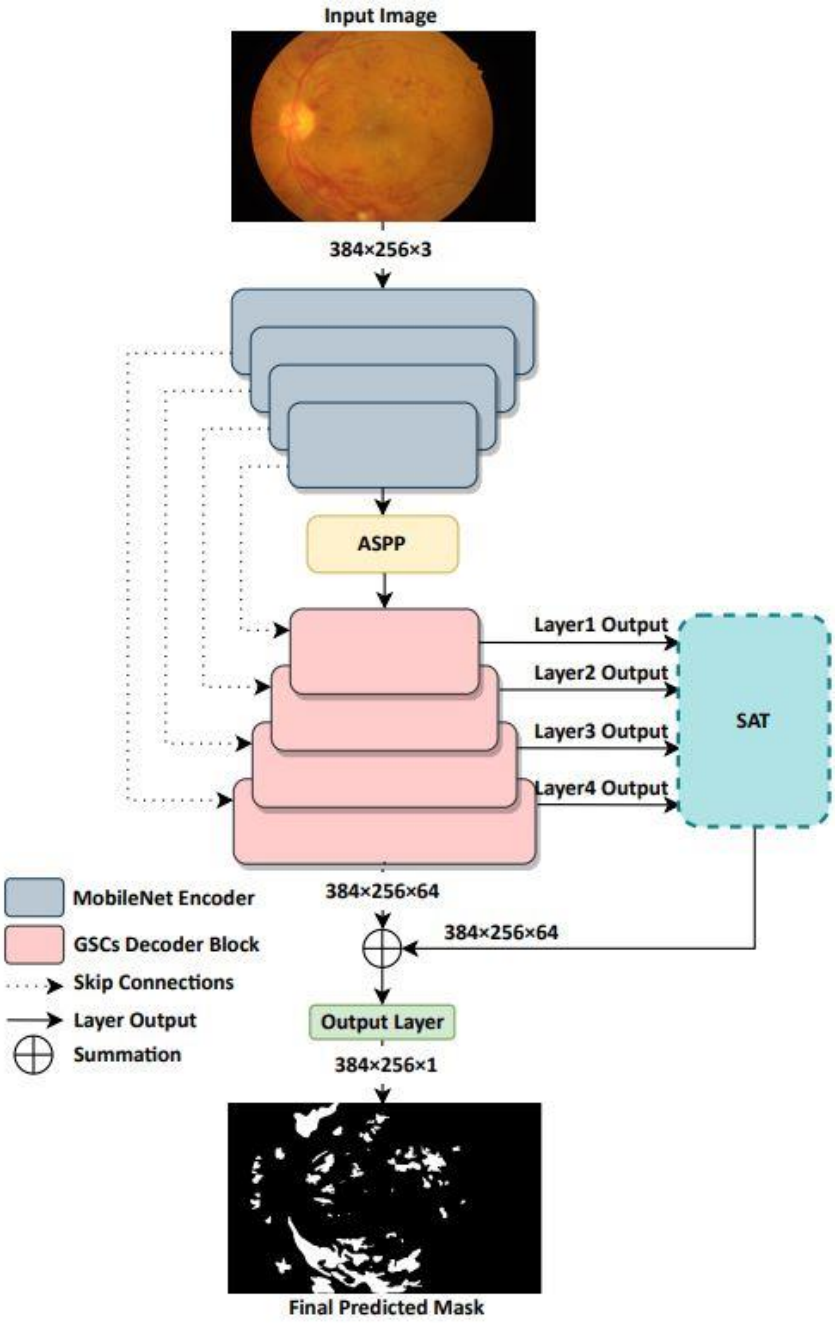


Figure 12. Architecture of the LezioSeg network for lesions segmentation in fundus images. [19]

3.4.1 Encoder Network

LezioSeg employs an ImageNet pre-trained MobileNet encoder as a backbone. This is because the MobileNet is a lightweight deep neural network with effective feature extraction capabilities and a cutting-edge foundation for many computer vision tasks. MobileNet uses depth-wise separable convolution. These types of CNNs are widely used because of the following two reasons: they have lesser number of parameters to adjust as compared to the standard CNN's, which reduces overfitting; and they are computationally

cheaper because of fewer computations. Also, MobileNet has two different global hyperparameters to reduce the computational cost-effectively [23].

The backbone includes four layers; it aims to encode the input eye fundus image and extract abstract information about retinal lesions at various levels of generality.

3.4.2 Neck of LezioSeg

The architecture includes an Atrous Spatial Pyramid Pooling (ASPP) module to aid in the extraction of multi-scale feature maps and to maximize the capture of contextual data of the small lesions. ASPP is a module for resampling a given feature layer at multiple rates. This is like probing the original image with multiple filters that have complementary effective fields of view, thus capturing objects as well as useful image context at multiple scales. Rather than resampling features, the mapping is implemented using multiple parallel atrous convolutional layers with different sampling rates. In this case it includes four parallel atrous convolutions [22].

The output of ASPP is the concatenated results of multi-scale feature maps, the decoder network follows the neck block of LezioSeg.

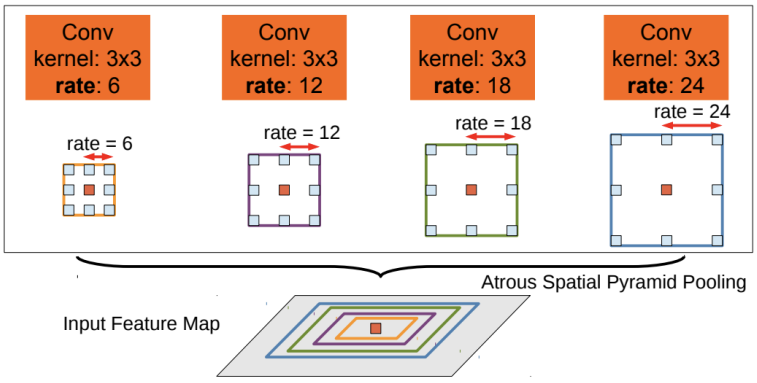


Figure 13. ASPP exploits multi-scale features by employing multiple parallel filters with different rates. [22]

3.4.3 Decoder Network

The decoder network includes four layers, SAT mechanism and the output layer that produces the final mask. Each decoder layer employs the GSCs mechanism followed by double convolution layers, batch normalization, and rectified linear unit activation function [19].

3.4.3.1 Gated Skip Connections (GSCs)

LezioSeg uses four GSCs blocks to boost feature map production and improve discrimination between the lesion and background pixels in retinal eye lesions segmentation.

Each GSCs decoder block receives feature maps from the corresponding MobileNet encoder block, which are concatenated with the feature maps produced by the previous block (either the ASPP neck block or a previous decoder block). After the concatenation, these features maps perform different operations which include: a convolution layer, multiplication, summation, and sigmoid activation.

After that, enhanced feature maps are fed into double convolution layers followed by batch normalization and rectified linear unit activation function. Finally, the output of each decoder layer is fed to the second multiscale block (SAT) [19].

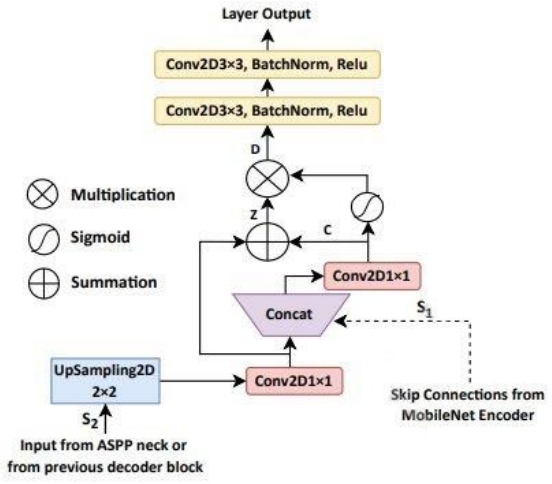


Figure 14. Architecture of GSC. [19]

3.4.3.2 Multi-Scale Attention (SAT) Mechanism

The multi-scale used to capture a wider range of relevant features with attention helps LezioSeg to maintain the multi-scale of each decoder block output to consider features from the four decoder blocks. In SAT, the four different copies of the features from the different stages of the decoder are collected to extract features and to reduce the dimension of features. Next, each scale size is upsampled to the original size of the input image using convolutional layers with different strides. After this, the four features maps perform different operations which include multiplication, summation, and sigmoid activation. And finally, in order to balance the SAT output, it is added to the final decoder network output, obtaining predicted mask for lesion segmentation [19].

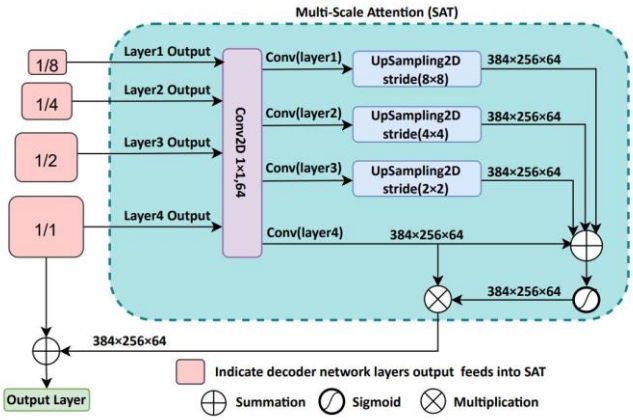


Figure 15. Architecture of SAT block. [19]

3.5 Diabetic Retinopathy Grading

The usual way of grading Diabetic Retinopathy from images is by training a deep learning model whose output is one of the DR severity grades. However, there is still a limitation in the performance of such approach, probably due to the difficulty of having a common procedure for labelling images by humans.

Recent works in the clinical field have established a rule-based procedure for determining the DR grade by observing the number of some types of lesions that appear in the eye fundus [24]. In that way, the ophthalmologists can have a common set of rules to assign de DR level to an image. In this work, we want to build a computer system that

follows this procedure based on lesions counting, instead of the classical approach. Therefore, in order to determine the degree of diabetic retinopathy in the patient, it is necessary to have an automatic method to count the number of lesions, both microaneurysms and haemorrhages, detected in the masks generated by the deep learning prediction models.

As far as the method of counting is concerned, it can generally be carried out in two different ways. The first is to count all the lesions present. The second method is to divide the image in half, both vertically and horizontally, resulting in four quadrants of equal dimensions. The two quadrants closest to the optic disc are called the nasal side and are recorded as quadrants 1 and 2. The two quadrants furthest from the optic disc are called the temporal side and are recorded as quadrants 3 and 4. It is important to emphasise the numbering because not all images show the optic disc on the right or left side of the image, as there are images of both eyes. In this second type of counting, only lesions present on the temporal side, i.e. in quadrants 3 and 4, are counted. This second counting approach is based on the observation that most lesions tend to be located in the temporal area of the eye, allowing a generalisation to be made by counting only this region.

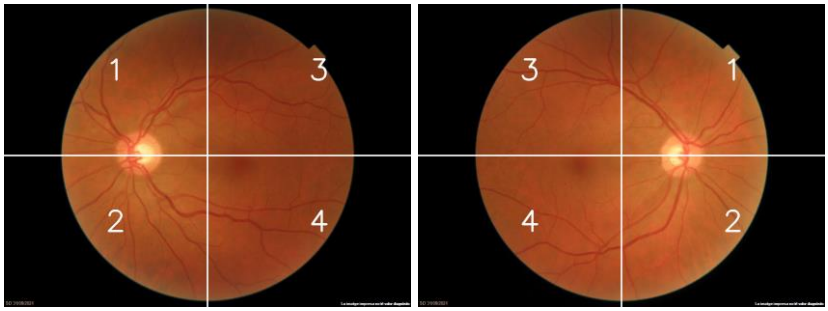


Figure 16. Fundus images of both eyes, showing the numbered quadrants.

Regarding the diagnosis of the DR grade from the number of lesions, there is currently a project called Messidor, which is created in order to evaluate automatic lesion segmentation and diabetic retinopathy grading methods [25]. This project has standardized the diagnosis of DR based on the number of lesions that may be present in an eye. It focuses on the quantity of microaneurysms and haemorrhages, and there are four different DR grades that follow the following rules:

- Grade 0: MA = 0 AND (HM = 0 OR HM = 1)
  No apparent retinopathy
- Grade 1: (0 < MA <= 5) AND HM = 0
  Mild diabetic retinopathy
- Grade 2: (5 < MA < 15) OR (0 < MA <= 5)
  Moderate diabetic retinopathy
- Grade 3: MA >= 15 OR HM >= 5
  Severe diabetic retinopathy

Occasionally, some datasets add a further grade called proliferative DR, as in the previous cases it was not. This grade considers the presence of neovascularization and/or haemorrhages.

3.6 Evaluation Metrics

This section describes the metrics used to analyse the various values obtained throughout the project. On the one hand, metrics will be obtained from the classification of

the resulting grades from the predictions, and on the other hand, metrics will be extracted from the testing phase of the new retrained LezioSeg models.

3.6.1 Metrics for Diabetic Retinopathy Grade Classification

A confusion matrix will be generated, which is a table showing the relationship between the predictions made by a model and the actual data provided by experts. From the values in the matrix, you can calculate various evaluation metrics, such as accuracy, recall, specificity, and F1-score. These metrics provide more detailed information about the performance of the model than simply looking at its overall accuracy. The metrics that were used are explained below.

Precision: refers to the proportion of instances that were correctly classified as positive out of all the instances that the model predicted as positive.

Recall: measures the proportion of positive instances that were correctly classified as positive.

F1-score: is the harmonic metric between accuracy and recall and can be useful when the balance between the two is important.

Support: is the number of true instances in each class. If this value is 0 it means that there are no true instances in that class.

All of the above metrics are assessed independently for each of the classes, i.e. for each of the predicted grades. In contrast, the metrics presented below are assessed on the entire dataset.

Accuracy : refers to the proportion of correct predictions in general.

Macro Average and **Weighted Average**: these metrics are averages of the individual metrics (accuracy, recall and F1-score) across all classes. In the case of macro-average, the metrics are averaged without taking into account class size imbalance. In the case of the weighted average, the metrics are weighted by the support of each class.

3.6.2 Metrics for Deep Learning Model Performance

The following metrics were used to assess the quality of the LezioSeg segmentation model.

Intersection-Over-Union (IOU)/Jaccard: it uses to check if there each landmark or abnormalities detector is over or under-segmentation of landmark or abnormalities regions. The IOU is the intersection ratio between the two masks concerning their union [26].

$$IOU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

F1-score: as it was mentioned before it stands for the harmonic mean of precision and recall. It can be expressed as follows [26]:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2}$$

Dice Coefficient: It is 2 times the overlap between the ground truth and the obtained mask, divided by the total number of pixels in both images. It can be expressed as follows [26]:

$$Dice\ Coefficient = \frac{2 * |A \cap B|}{|A \cup B|} \tag{3}$$

Area Under Precision-Recall curve (AUPR): This curve shows the trade-off between precision and recalls for different threshold values. The high area under the curve represents high recall and precision, indicating good performance. It is known to be a realistic measure for lesion segmentation performance like lesions.

3.7 Development Environment

The project was completed utilizing Python, due to its wide range of libraries that facilitated the execution of various tasks. Some of the most commonly used libraries are described below.

For the development of the lesion prediction part, the TensorFlow library was chosen [27]. This library allows the creation and training of automatic learning models. This allowed us to carry out the entire process of creating and assembling the LezioSeg neural network, followed by the training, testing and inference phases. This library was also used on various occasions to process the masks during the prediction process and also facilitated the process of creating data sets compatible with the neural network. Furthermore, given the lack of experience in the field of deep learning, extensive use was made of the tools provided by Keras [28]. Keras is a library that can run on top of TensorFlow. It offers a more straightforward and high-level set of functions, making development easier.

To analyse the masks obtained from the predictions and the ground truth, the OpenCV[2] and PIL[3] libraries were used [29], [30]**,** both of which allow image processing, including the manipulation of pixels or the detection of objects in the images.

The obtained data from the masks was stored in CSV files, and subsequently transferred into two-dimensional data structures, known as DataFrames, for processing and manipulation in Python. The operations were conducted utilizing the Pandas library, which encompasses all indispensable tools.

The analysis of acquired results utilised the Scikit-learn library, more specifically the sklearn.metrics module which offers instruments for validating and appraising machine learning models [31]. Additionally, the manipulation of matrices was facilitated through the incorporation of NumPy library [32].

Jupyter Notebook was utilised to edit the code, as the ability to execute independent blocks of code without having to rerun the entire programme brings a degree of flexibility to the development process. This feature makes debugging easier at later stages.

In relation to the use of the models, computer equipment on the URV's ITAKA laboratory was used, which is equipped with a GTX 1070 GPU.

[2] Open Source Computer Vision Library

[3] Python Imaging Library

4 Design and Implementation

This project continues the work of Salem's PhD thesis at the research group ITAKA in a project about Diabetic Retinopathy diagnoses that started 10 years ago. Therefore, the design and implementation have required to acquire the knowledge of the project framework and tools already developed. The work workflow that has been carried out is detailed in *Figure 17*.

First, tasks related to the analysis of the already predicted lesion masks for IDRiD dataset were carried out, so the first tasks performed were related to lesion counting and subsequent diagnosis, based on the previous count. Once the DR scores were obtained, they were compared with the scores diagnosed by the experts or with the scores obtained from the ground truth masks.

Next, other datasets were obtained to validate the results of IDRiD. Then, the different neural networks that were already trained for each type of lesion were used, and an inference phase was carried out on them with new datasets. Then, once the new predicted masks for this new dataset had been obtained, the previously mentioned steps related to the analysis were carried out again.

Finally, the deep learning LezioSeg was retrained by creating a new dataset with images from the Sant Joan de Reus Hospital, which was used in the training phase and for the subsequent testing of the new LezioSeg weights.

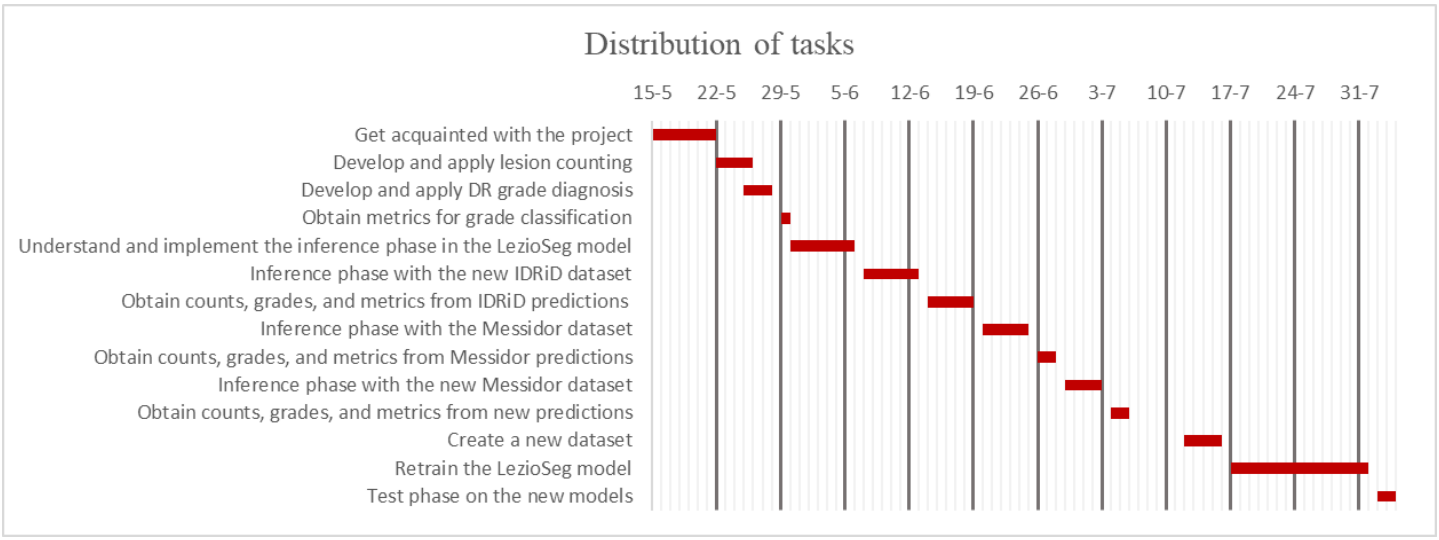


Figure 17. Gantt chart of project's development.

The final list of tasks designed and implemented is the following one, and they are explained in the rest of sections of this chapter.

- Lesions Counting from masks.
- Diagnosis of the Diabetic Retinopathy graded from the lesions counts.
- Evaluation of the predicted gradings.
- LezioSeg inference with new testing data.
- LezioSeg retraining with new data.

4.1 Lesions Counting

Regarding the counting of the masks, as defined in *3.5*, it can be done in two ways: counting all the lesions or counting only those present on the temporal side of the eye, being quadrants 3 and 4.

For the second method mentioned it is necessary to know the position of the optic disc. As not all datasets have this information, two divergent functions have been created, one that performs both types of counting and one that only counts the whole image.

When both types of counting are performed, the counting of the total number of lesions consists only of obtaining the total number of contours in the image. For the other type of counting, the image is divided into four equal quadrants, and the number of contours present in each quadrant is determined. Subsequently, the position of the optic disc is determined, because if it is on the right side of the image, the values of quadrants 1 and 2 must be permuted by the values of quadrants 3 and 4, so that the lesions furthest from the optic disc are found in the latter two quadrants. In the case of only performing the total count, it is carried out as explained above.

This counting process can be performed on masks with predicted lesions and on masks with lesions marked by professionals. All of this data is stored in CSV files for later processing, where each row refers to a fundus image and each column represents a count performed.

4.2 Diagnosing Diabetic Retinopathy Grade

When determining the degree of diabetic retinopathy based on the number of lesions, is available the following information: the total and/or temporal side count of microaneurysms and haemorrhages of predicted masks and/or ground truth masks.

Table 1. Summary of possible mask types and counts.

		Type of Mask	
		Predicted	Ground Truth
Type of Counting	Whole Image	Pred_Total	GT_Total
	Quadrants 3 and 4	Pred_q34	GT_q34

To diagnose the degree of DR, the CSV files containing the lesions must first be manipulated. First, the counts for microaneurysms and haemorrhages are merged to obtain the number of both lesions for each image. It should be noted that this merging is done so that the name of each image is the same in both files. Once all the data are available, the rules mentioned in *3.5* are applied to them, assigning the corresponding grade to each image on the basis of the number of MA and HM. A new column is created with the corresponding grade. If it does not fit any of the four rules, it is recorded as grade -1 for that image.

Again, this process can be carried out on the count of masks with lesions that have been predicted, and on the count of masks with lesions that have been marked by professionals.

4.3 Evaluating Grade Predictions

In this stage, the grades obtained by predicting the lesions are compared with the grades obtained with the masks marked by the experts, or in other cases where the experts have directly diagnosed the grade, so that it was not necessary to carry out the previous stages.

First, a confusion matrix is created to visualise the accuracy of the predictions. This matrix allows us to systematically compare the data obtained by the predictions with the data provided by experts in the field of study. Although there are already implemented functions, a new function was created to check for null values, i.e. -1. The function consists of creating a matrix, and each predicted degree with its corresponding degree ground truth function as coordinates in this matrix, this duple adds a unit in the "position" that it indicates. In this way, in a confusion matrix, the predicted degrees can act as the vertical axis and the ground truth degrees as the horizontal axis, or vice versa.

In addition, a text report is generated showing the main classification metrics, such as precision, recall and F1-score, for each class (DR grade) in the dataset. It also provides the average precision, recall and F1-score for the entire dataset. This requires two data sets: a set of true values and a set of predicted values.

4.4 Inference Phase

In this phase, LezioSeg models were used for each lesion type, which had been previously trained on a different dataset. In addition, it was available a set of functions used for processing both the dataset used for inference and the masks used. These ready-to-use tools have greatly simplified the process.

The procedure for predicting the masks used to identify the lesions is as follows. First, the neural network is configured, a procedure that involves the creation of a model with the specific architecture for LezioSeg, followed by the loading of the weights previously trained and stored. Finally, a designated function is invoked for the prediction of the different lesions, generating, and storing the corresponding masks. It should be noted that this inference phase has been carefully applied both for the segmentation of both essential lesions and the optic disc. The latter is essential in the context where both counting modalities mentioned above are desired.

4.5 Retraining LezioSeg Model

In this phase, a training and fine-tuning process of the LezioSeg neural network was carried out. Two essential elements are required for this: the acquisition of a new data set for the next training and future testing, as well as the availability of the deep learning network architecture to be used in the training process.

4.5.1 Dataset Preparation and Distribution

Regarding the issue of the dataset, it is necessary that this dataset is composed of a new set of images that are different from those used in the previous training. This will allow any changes to be evaluated during the new retraining process. This set should contain at least retinal fundus images and corresponding lesion masks marked by specialists. This arrangement allows for prediction and comparison during the training process by contrasting the predicted masks with the ground truth values.

Another important aspect in the construction of the dataset is its distribution. In this case, 80% of the images and masks were reserved for the training phase, while the

remaining 20% were used for testing. Within the group reserved for training, 80% of these samples were used for the training process itself and the remaining 20% were used for the validation phase. Care was taken to balance the distribution of the different image types across the three subsets mentioned above, in order to include images with different proportions of lesions.

Once the dataset was split and structured, it was ready for use. Two datasets were created for the development of the project. These datasets differ in that the first consists of complete images, while the second divides the images into four equal parts. The purpose of this division is to investigate whether there is an improvement over the first set. In both cases, a data augmentation process was applied to increase the breadth of the dataset.

4.5.2 LezioSeg Retraining Process

Once the different datasets are available, the LezioSeg retraining process is carried out, since the neural network will be equipped with this architecture. This procedure is largely based on the work previously carried out by Salem, using the parameters previously configured during the training carried out by himself, as well as various functions implemented for different tasks. These functions include mask processing and visualisation, as well as obtaining metrics and visualising the training process.

Some of the reused parameters include elements such as the number of epochs. Binary cross-entropy was used for the loss function, and the Adam optimization algorithm was implemented as optimizer.

In each epoch of training, the images are passed with a certain batch size, which indicates the number of images that are passed through the model at the same time. In general, for training, one image at a time is passed through the model, so the size is 1.

Additionally at the end of each training epoch, the validation phase is performed, during which several metrics are collected. These metrics allow us to assess whether the model has improved from its previous state. If so, the weights of this new model are stored.

Another aspect to consider during training is the use of the fine-tuning process. This refers to the choice between starting training from scratch or using a previously trained model. Both approaches are carried out in order to make comparisons and evaluate the results obtained in each case.

4.6 Evaluation and Testing of Model Performance

The final phase of the project aims to evaluate the ability of the models obtained from the different training processes to make accurate predictions on new, previously unobserved data. At this stage, the tools previously implemented by Salem are again used.

In this context, the appropriate dataset is used, as described in the previous section. On this dataset, predictions of both types of lesions are made using the newly generated models. These predicted masks are compared to the masks with ground truth values, allowing the evaluation of essential metrics for the evaluation of these new models. Unlike the training phase, no data augmentation is required at this stage, as this process is more akin to a practical exercise in reality.

The completion of this evaluation phase will provide an informed analysis of the models' capability and performance, giving an objective view of their effectiveness in making accurate predictions in previously unknown contexts.

5 Results

This section provides a comprehensive breakdown of the implemented workflow. A description of the data used at each stage will be provided, along with the specific functions that were applied. In addition, the results obtained during the process are presented and the decisions made based on these results are discussed.

5.1 Diabetic Retinopathy Diagnosis and Analysis Using IDRiD Dataset

The starting point of the project is the pixel-annotated dataset extracted from the IDRiD database and used by Salem during the evaluation phase of the LezioSeg model. This dataset consists of 27 images, together with the predicted masks and the ground truth. This database was used for the detailed analysis described below.

First, microaneurysms and haemorrhages present in each mask were counted. The availability of the optic disc masks allowed two types of counts to be made: total and by quadrants. Once the number of lesions present was obtained, Messidor's diagnostic rules were applied. This allowed the degree of retinopathy to be determined for both the predicted and expert graded masks. Importantly, two diagnoses were derived for each mask type: one based on the total count, and one based on third and fourth quadrants. This resulted in a total of four types of diagnoses, as shown in *Table 1*Table 1. Summary of possible mask types and counts.. As a result, five grades of diabetic retinopathy were available for each image: two derived from the predictions, two based on the ground truth masks, and one provided by a specialist.

Finally, several comparisons were made between these grades. The grades obtained from the predictions were compared with those obtained from the ground truth masks, taking into account both the total count and the counts in quadrants 3 and 4. In addition, the grades from the predictions were compared with the diagnostic grades determined by a specialist. In this way, four different scores were presented.

Table 2. Confusion matrix of predicted diabetic retinopathy grades versus the ground truth grades, using the total count method.

	GT_Total_0	GT_Total _1	GT_Total _2	GT_Total _3
Pred_Total_0	0	0	0	0
Pred_Total _1	0	0	0	0
Pred_Total _2	0	0	4	0
Pred_Total _3	0	0	0	23

Table 3. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 2	1.00	1.00	1.00	4
Class 3	1.00	1.00	1.00	23
Accuracy	-	-	1.00	27
Macro Average	1.00	1.00	1.00	27

Weighted Average	1.00	1.00	1.00	27

Table 4. Confusion matrix of predicted diabetic retinopathy grades versus the ground truth grades, using the counting method of quadrants 3 and 4.

	GT_q34_0	GT_q34_1	GT_q34_2	GT_q34_3
Pred_q34_0	0	0	0	0
Pred_q34_1	0	0	0	0
Pred_q34_2	0	0	7	3
Pred_q34_3	0	0	4	12

Table 5. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 2	0.64	0.70	0.67	10
Class 3	0.81	0.76	0.79	17
Accuracy	-	-	0.74	27
Macro Average	0.72	0.73	0.73	27
Weighted Average	0.75	0.74	0.74	27

Table 6. Confusion matrix of predicted diabetic retinopathy grades versus the expert diagnosis, using the total count method.

	Expert_0	Expert _1	Expert _2	Expert _3
Pred_Total_0	0	0	0	0
Pred_Total _1	0	0	0	0
Pred_Total _2	1	1	0	2
Pred_Total _3	0	0	0	23

Table 7. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.00	0.00	0.00	0
Class 1	0.00	0.00	0.00	0
Class 2	0.00	0.00	0.00	4
Class 3	0.92	1.00	0.96	23
Accuracy	-	-	0.85	27

Macro Average	0.23	0.25	0.24	27
Weighted Average	0.78	0.85	0.82	27

Table 8. Confusion matrix of predicted diabetic retinopathy grades versus the expert diagnosis, using the counting method of quadrants 3 and 4.

	Expert_0	Expert _1	Expert _2	Expert _3
Pred_q34_0	0	0	0	0
Pred_q34_1	0	0	0	0
Pred_q34_2	1	1	0	8
Pred_q34_3	0	0	0	17

Table 9. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.00	0.00	0.00	0
Class 1	0.00	0.00	0.00	0
Class 2	0.00	0.00	0.00	10
Class 3	0.68	1	0.81	17
Accuracy	-	-	0.63	27
Macro Average	0.17	0.25	0.20	27
Weighted Average	0.43	0.63	0.51	27

The performance indicators are excellent when comparing the ground truth given at IDRiD with the predictions of our system. However, when it was checked against the medical experts in our team, the quality was reduced.

Due to the small size of this dataset and its bias towards the worst class, conclusions have low significance. Therefore, more images were searched for a performing a more complete testing.

5.2 Inference and Analysis of the New IDRiD Dataset

Due to the mentioned limitations of the initial testing set, it was decided to include a new dataset. This new approach used another dataset from the IDRiD database, consisting of 516 fundus images. This dataset also contained information on retinopathy grades but lacked any associated masks. Therefore, in this new phase of the project, it was necessary to first perform the inference process using the already trained LezioSeg models. This strategy allowed us to obtain predicted masks for microaneurysms, haemorrhages and the disc.

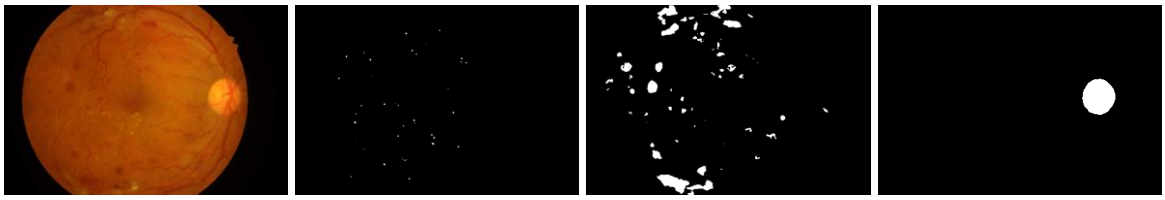


Figure 18. Retinal fundus image and its respective predicted masks, from left to right: microaneurysms, haemorrhages and optic disc.

All the steps performed in the previous phase of the project were then replicated. As a result of this replication, two diagnoses were derived based on lesion counts according to the predicted masks, either a global count or a count of quadrants 3 and 4. After obtaining these grades, a comparison was made with the pre-existing grades in the original dataset. It is important to note that this dataset categorised DR grades on a scale of 0 to 4, in contrast to the classification used in this project which is limited to grade 3. After discussion with experts in the field, it was agreed that images classified as grade 4 should be adjusted to grade 3. This was based on the consideration that grade 4 in this dataset only indicates the presence of neovascularisation in the eye. As a result, two sets of comparisons were generated, depending on the type of counting performed.

Table 10. Confusion matrix of predicted diabetic retinopathy grades versus grades diagnose by the dataset, using the total count method.

	IDRiD_0	IDRiD _1	IDRiD _2	IDRiD _3
Pred_Total_0	17	0	0	0
Pred_Total _1	46	5	4	0
Pred_Total _2	98	19	89	32
Pred_Total _3	7	1	75	123

Table 11. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.10	1.00	0.18	17
Class 1	0.20	0.09	0.13	55
Class 2	0.53	0.37	0.44	238
Class 3	0.79	0.60	0.69	206
Accuracy	-	-	0.45	516
Macro Average	0.41	0.52	0.36	516
Weighted Average	0.59	0.45	0.49	516

Table 12. Confusion matrix of predicted diabetic retinopathy grades versus grades diagnose by the dataset, using the counting method of quadrants 3 and 4.

	IDRiD_0	IDRiD _1	IDRiD _2	IDRiD _3
Pred_q34_0	54	2	5	3
Pred_q34_1	51	6	10	0
Pred_q34_2	57	17	102	45
Pred_q34_3	6	0	51	107

Table 13. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.32	0.84	0.47	64
Class 1	0.24	0.09	0.13	67
Class 2	0.61	0.46	0.52	221
Class 3	0.69	0.65	0.67	154
Accuracy	-	-	0.52	516
Macro Average	0.46	0.51	0.45	516
Weighted Average	0.55	0.52	0.51	516

After observing that the results, we can see that the performance is poor, with an overall accuracy of 0.52, and low scores for average of F1 score as well. This means that the LezioSeg model is not able to be used with the lesion counting procedure for DR grading in this dataset.

5.3 Inference and Analysis of the Messidor Dataset

To study if the problem is with the type of images of IDRID dataset, we took a different dataset, with images captured by different types of machines and with different population: Messidor. This new set consisted of 1200 images that had already been divided into the 4 degrees of DR that can be represented. After segmenting the lesions and obtaining the diagnoses, the following metrics were obtained.

Table 14. Confusion matrix of predicted diabetic retinopathy grades versus the expert diagnosis, using the total count method.

	Expert_0	Expert _1	Expert _2	Expert _3
Pred_Total_0	226	11	8	3
Pred_Total _1	229	79	32	7
Pred_Total _2	151	88	112	163
Pred_Total _3	3	1	1	86

Table 15. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.37	0.91	0.53	248
Class 1	0.44	0.23	0.30	347
Class 2	0.73	0.22	0.34	514
Class 3	0.33	0.95	0.49	91
Accuracy	-	-	0.42	1200
Macro Average	0.47	0.58	0.41	1200
Weighted Average	0.54	0.42	0.38	1200

Table 16. Confusion matrix of predicted diabetic retinopathy grades versus the expert diagnosis, using the counting method of quadrants 3 and 4.

	Expert_0	Expert _1	Expert _2	Expert _3
Pred_q34_0	325	30	27	11
Pred_q34_1	175	80	40	17
Pred_q34_2	107	69	86	174
Pred_q34_3	2	0	0	57

Table 17. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.53	0.83	0.65	393
Class 1	0.45	0.26	0.33	312
Class 2	0.56	0.20	0.29	436
Class 3	0.22	0.97	0.36	59
Accuracy	-	-	0.46	1200
Macro Average	0.44	0.56	0.41	1200
Weighted Average	0.51	0.46	0.42	1200

It should be noted that the results are quite poor, as none of the metrics represented have high values. Looking at the confusion matrices, it can be observed that in general the model tends to predict a higher degree than the corresponding one, this is due to the fact that, when it comes to recognising lesions, it recognises more than those that actually exist.

After observing these results with a larger dataset, the main conclusion is that models are not properly trained to detect such lesions. But by reviewing the process used to predict

such images was found a factor to consider. This was that the input size that the images must have in the model is 768 × 512, while Messidor's images have a size of 640 × 640. So, when the dataset is created, the image is resized to fit the required dimension. This causes a deformation of the image, which can lead to erroneous predictions being made.

To try to mediate this problem, a function has been created that resizes the image taking into account the width and height required in the neural network input and maintaining the aspect ratio. In this case, as it is a square and the input is a rectangle, it does not fit perfectly, so this same function adds equal black bars on each side.

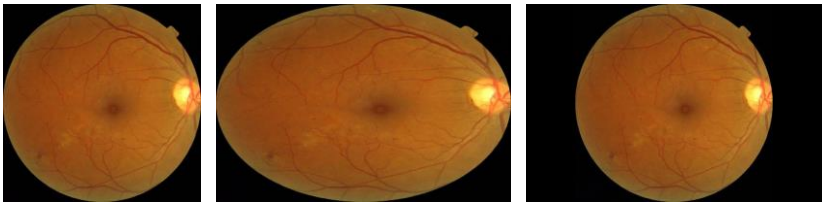


Figure 19. From left to right: original image, image resized to LezioSeg input size and image with LezioSeg input size but not distorted.

After making these modifications, which allow the original image and the required size to be maintained, the prediction of all the images was carried out again, obtaining the following results.[33]

Table 18. Confusion matrix of predicted diabetic retinopathy grades versus the expert diagnosis, using the total count method.

	Expert_0	Expert _1	Expert _2	Expert _3
Pred_Total_0	202	15	8	3
Pred_Total _1	194	59	38	10
Pred_Total _2	208	104	105	179
Pred_Total _3	5	1	2	67

Table 19. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.33	0.89	0.48	228
Class 1	0.33	0.20	0.25	301
Class 2	0.69	0.18	0.28	596
Class 3	0.26	0.89	0.40	75
Accuracy	-	-	0.36	1200
Macro Average	0.40	0.54	0.35	1200
Weighted Average	0.50	0.36	0.32	1200

Table 20. Confusion matrix of predicted diabetic retinopathy grades versus the expert diagnosis, using the counting method of quadrants 3 and 4.

	Expert_0	Expert _1	Expert _2	Expert _3
Pred_q34_0	319	41	28	16
Pred_q34_1	156	65	48	18
Pred_q34_2	132	73	74	160
Pred_q34_3	2	0	3	65

Table 21. Report of the main classification metrics, using the above classification.

	Precision	Recall	F1-score	Support
Class 0	0.52	0.79	0.63	404
Class 1	0.36	0.23	0.28	287
Class 2	0.48	0.17	0.25	439
Class 3	0.25	0.93	0.40	70
Accuracy	-	-	0.44	1200
Macro Average	0.41	0.53	0.39	1200
Weighted Average	0.45	0.44	0.39	1200

After these results it can be seen that they even get worse, since, as shown in the confusion matrix, in each of the grades the number of correct predictions is lower than in the previous case.

As previously mentioned, after ruling out that the poor results obtained in the Messidor dataset were due to the images being deformed, it is clear that the neural network is not well trained. Subsequent consultations with experts specializing in the detection of these types of lesions have revealed that certain masks within the IDRiD dataset, which served as the training data for the neural network, are inaccurate. These masks erroneously identify lesions that do not exist in the corresponding eye.

It should also be noted that, apart from being few images, many of them have a high degree of diabetic retinopathy, so the diversity of grades and thus the number of lesions is also low. This argument and the fact that there is a marking of erroneous ground truth would explain why the predicted grade is often higher than the real one, as the number of false positives increases, due to the way the network is trained.

It's crucial to remember that it is necessary to train a model that has a high accuracy rate. One of the main goals is to detect as accurately as possible the number of microaneurysms present. This is because in those cases where patients present few microaneurysms, they are in the early stages of the disease, and it is crucial to treat as soon as it is detected.

5.4 Retraining and Testing LezioSeg

In view of these disappointing results, it would be necessary to try again to train the neural networks. For the new training of LezioSeg, it was decided to use a dataset consisting of images from the Sant Joan de Reus Hospital, annotated by an expert. In this case, it was necessary to perform some preprocessing to obtain the masks, since the annotated lesions were stored in CSV files instead of actual image masks. Consequently, these files, which contained all the necessary information, first had to be converted into readable images. Once a suitable set of images and masks was obtained, the appropriate splits were performed to generate the three required datasets.

Different trainings are performed depending on the dataset configurations used. Some of the variations considered include whether the images in the training set are complete or smaller fragments obtained from the whole image. Another variable is the generation of multiple replicates of the dataset to perform data augmentation, where certain parameters are varied to create a large and diverse set of images, sets of 8 and 12 replicates were created. In addition, the option of training from the weights trained in previous models or from scratch is considered. A summary of all training performed is shown in the following table.

Table 22. Summary of the different dataset configurations used to retrain the LezioSeg model.

	Whole Images		Split Images	
From IDRiD	8 repetitions	12 repetitions	8 repetitions	12 repetitions
From Scratch	8 repetitions	12 repetitions	8 repetitions	12 repetitions

8 different training sessions were performed for each lesion, with 30 epochs in each and with the hyperparameters defined in *4.5.2.* Finally, after having stored the weights of each model obtained by training with various dataset configurations, covering both types of lesions, a testing phase was carried out with each of the resulting models, from which will be obtained the metrics detailed in *Section 3.6.2*.

Metrics obtained for microaneurysm segmentation.

Table 23. IOU obtained in the test phase, for each of the models trained for the segmentation of microaneurysms.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.03	0.2	0.01	0.22
From Scratch	0.02	0.15	0.03	0.22

Table 24. F1-score obtained in the test phase, for each of the models trained for the segmentation of microaneurysms.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.06	0.31	0.03	0.34
From Scratch	0.17	0.25	0.17	0.33

Table 25. Dice coefficient obtained in the test phase, for each of the models trained for the segmentation of microaneurysms.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.06	0.31	0.03	0.34
From Scratch	0.05	0.25	0.05	0.33

Table 26. AUPR obtained in the test phase, for each of the models trained for the segmentation of microaneurysms.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.26	0.4	0.31	0.41
From Scratch	0.17	0.44	0.17	0.4

Metrics obtained for haemorrhage segmentation.

Table 27. IOU obtained in the test phase, for each of the models trained for the segmentation of haemorrhages.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.3	0.45	0.34	0.44
From Scratch	0.28	0.0006	0.23	0.38

Table 28. F1-score obtained in the test phase, for each of the models trained for the segmentation of haemorrhages.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.44	0.6	0.49	0.58
From Scratch	0.42	0.001	0.35	0.51

Table 29. Dice coefficient obtained in the test phase, for each of the models trained for the segmentation of haemorrhages.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.44	0.6	0.49	0.58
From Scratch	0.42	0.001	0.35	0.51

Table 30. AUPR obtained in the test phase, for each of the models trained for the segmentation of haemorrhages.

	8 Repetitions		12 Repetitions	
	Whole Images	Split Images	Whole Images	Split Images
From IDRiD	0.49	0.64	0.55	0.63
From Scratch	0.49	0.25	0.44	0.57

None of the models created with the new training strategies achieved outstanding values in the evaluated metrics. However, it is important to note that only one training process was carried out due to time limitations. Training with the exploration and variation of different parameters should be done and probably will lead to more encouraging results.

Regarding the results, it is generally noted that the models that underwent fine-tuning produced more favourable metrics in comparison to the models trained from scratch. Moreover, when models were trained with image fragments instead of using full images, slightly better results were also achieved. In contrast, the difference in results between datasets with 8 repetitions and those with 12 repetitions was less significant, with values frequently being similar.

It is noteworthy that there are signs of inadequate performance in the haemorrhage segmentation model trained from scratch, when utilising a dataset composed of 8 repetitions of fragmented images. In this configuration, notoriously low metrics were observed compared to other models.

An important conclusion to consider is that, in general, more satisfactory results were obtained in the segmentation of haemorrhages than in the segmentation of microaneurysms. This finding is consistent with the difference in size of the datasets used, with the dataset for haemorrhage segmentation being slightly larger than that for microaneurysm segmentation. However, it is important to note that this disparity in performance can also be attributed to the nature of the lesions themselves. Haemorrhages tend to be more visible and detectable on images compared to microaneurysms, which tend to present as small, subtle dots.

In order to evaluate the metrics obtained in this retraining, a comparison was made with the metrics obtained by Salem during his training with the IDRiD dataset.

Table 31. Metrics obtained by LezioSeg training performed by Salem with the IDRiD dataset, compared with the best metrics obtained in this project. Using for microaneurysms the fine-tuned model trained with a dataset of 12 repetitions of split images and for haemorrhages the fine-tuned model trained with a dataset of 8 repetitions of split images.

	Salem metrics			Best retrain metrics		
	IOU	F1-score	AUPR	IOU	F1-score	AUPR
Microaneurysms	0.6	0.34	0.38	0.22	0.34	0.41
Haemorrhages	0.7	0.59	0.67	0.45	0.6	0.64

In order to evaluate the metrics obtained in this retraining, a comparison was made with the metrics obtained by Salem during his training with the IDRiD dataset. Overall, the results obtained in this retraining showed a slight decrease in metrics compared to Salem's results. However, it is notable that this difference was not significantly large, and in some cases, even higher metrics were achieved. Although the project results are lower, this

overall comparison with another training process suggests the possibility of carrying out an additional inference phase using the previous datasets. This would allow an assessment of whether the predictions of diabetic retinopathy grades are more accurate and therefore provide a more complete picture of the model's capability.

6 Discussion and Conclusion

In the Bachelor Project the possibility of grading the severity of Diabetic Retinopathy by means of the detection of the number of lesions in eye fundus images has been studied.

The work was based on a deep learning architecture called LezioSeg, which has been studied in depth in order to be able not only to use it for inference but also for the retraining of its parameters using a new images dataset.

Being a work in the frame of research, it is possible to end in a situation with not good results. After the extensive testing phase carried out on with different data and with different variations, the initial hypothesis of using LezioSeg model to make a good prediction of DR has not been confirmed. In general, it can be concluded that the results obtained did not reach optimal levels, or at least not better.

These findings indicate that additional research and enhancements to the training process are required. A first step in this direction has been done by using a dataset created in a controlled environment at Hospital Sant Joan de Reus. However, we have not been able to construct a new version good enough. With some more time, various parameters and strategies should be explored to enhance the performance of LezioSeg models in eye lesion segmentation.

The results obtained in this study suggest several probable causes for the sub-optimal performance of the LezioSeg models. It is essential to remember that this training represents a starting point, and that further research will be needed to improve the results. Some possible factors that may have led to the suboptimal findings are:

- Increase the size of the training dataset, despite being larger than that of IDRiD, it is possible that the amount of data is still not sufficient to achieve optimal performance.
- To consider are minimizing black areas in the training images, which do not provide any information and may hamper the model performance.

In terms of the counting methods used, better outcomes were generally observed when diagnosing exclusively quadrants 3 and 4. It should be noted that this does not imply that this method is more accurate, as it may be the result of over-predicting lesions and improving the results by reducing the number of quadrants to consider. However, the experts have decided to take the total count into account on future occasions. This simplifies the diagnostic process by removing the need to detect the optic disc and is based on an overall lesion count, rather than dividing it into quadrants.

In the future, if acceptable metrics are achieved, it will be possible to predict diagnoses with new sets of images taken at the hospital. This would allow a complete diagnostic process for diabetic retinopathy to be carried out and the results to be compared with expert diagnoses. Nevertheless, it is worth noting that the segmentations required for this type of diagnosis must be highly accurate and often complex, as lesions can be difficult to detect. In addition, alterations in the brightness, texture or colour of the images may impact the precision of the segmentation. Therefore, these predictions are sensitive to the specific characteristics of the datasets used.

In conclusion, diabetic retinopathy diagnosis based on eye lesion segmentation is a promising approach that requires further development and refinement.

This project has presented an opportunity to enhance collaborative skills and tackle challenges in a multidisciplinary and multicultural teamwork environment. Moreover, it has also provided a valuable introduction to the field of deep learning, which has enabled me the acquisition of significant knowledge on machine learning, computer vision and artificial intelligence, as well as useful engineering skills.

7 References

[1] N. Cheung, P. Mitchell, and T. Y. Wong, “Diabetic retinopathy,” *Lancet*, vol. 376, no. 9735, pp. 124–136, 2010, doi: 10.1016/S0140-6736(09)62124-3.

[2] P. Romero-Aroca, R. Navarro-Gil, A. Valls-Mateu, R. Sagarra-Alamo, A. Moreno-Ribas, and N. Soler, “Differences in incidence of diabetic retinopathy between type 1 and 2 diabetes mellitus: a nine-year follow-up study,” *Br J Ophthalmol*, vol. 101, no. 10, p. 1346, Oct. 2017, doi: 10.1136/BJOPHTHALMOL-2016-310063.

[3] “Diabetic Retinopathy Honolulu | Diabetic Eye Disease Kailua-Kona.” https://www.bennetteyeinstitute.com/retina-honolulu/diabetic-retinopathy/ (accessed Sep. 04, 2023).

[4] Z. L. Teo *et al.*, “Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis,” *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, Nov. 2021, doi: 10.1016/J.OPHTHA.2021.04.027.

[5] P. Romero-Aroca, “Ocular Complications of Diabetes and Therapeutic Approaches,” *Journal of Clinical Medicine 2022, Vol. 11, Page 5170*, vol. 11, no. 17, p. 5170, Sep. 2022, doi: 10.3390/JCM11175170.

[6] G. Kumar, S. Chatterjee, and C. Chattopadhyay, “DRISTI: a hybrid deep neural network for diabetic retinopathy diagnosis,” *Signal Image Video Process*, vol. 15, no. 8, pp. 1679–1686, Nov. 2021, doi: 10.1007/S11760-021-01904-7.

[7] A. W. Stitt *et al.*, “The progress in understanding and treatment of diabetic retinopathy,” *Prog Retin Eye Res*, vol. 51, pp. 156–186, Mar. 2016, doi: 10.1016/J.PRETEYERES.2015.08.001.

[8] R. Srivastava, L. Duan, D. W. K. Wong, J. Liu, and T. Y. Wong, “Detecting retinal microaneurysms and hemorrhages with robustness to the presence of blood vessels,” *Comput Methods Programs Biomed*, vol. 138, pp. 83–91, Jan. 2017, doi: 10.1016/J.CMPB.2016.10.017.

[9] R. J. Chalakkal, W. H. Abdulla, and S. C. Hong, “Fundus retinal image analyses for screening and diagnosing diabetic retinopathy, macular edema, and glaucoma disorders,” *Diabetes and Fundus OCT*, pp. 59–111, Jan. 2020, doi: 10.1016/B978-0-12-817440-1.00003-6.

[10] B. Zhang, X. Wu, J. You, Q. Li, and F. Karray, “Detection of microaneurysms using multi-scale correlation coefficients,” *Pattern Recognit*, vol. 43, no. 6, pp. 2237–2248, Jun. 2010, doi: 10.1016/J.PATCOG.2009.12.017.

[11] M. W. Nadeem, H. G. Goh, M. Hussain, S. Y. Liew, I. Andonovic, and M. A. Khan, “Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions,” *Sensors (Basel)*, vol. 22, no. 18, Sep. 2022, doi: 10.3390/S22186780.

[12] H. Jiang *et al.*, “A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation,” *Comput Biol Med*, vol. 157, p. 106726, May 2023, doi: 10.1016/J.COMPBIOMED.2023.106726.

[13] P. Ongsulee, “Artificial intelligence, machine learning and deep learning,” *International Conference on ICT and Knowledge Engineering*, pp. 1–6, Jan. 2018, doi: 10.1109/ICTKE.2017.8259629.

[14] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[15] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” *Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017*, vol. 2018-January, pp. 1–6, Mar. 2018, doi: 10.1109/ICENGTECHNOL.2017.8308186.

[16] “What are Convolutional Neural Networks? | IBM.” https://www.ibm.com/topics/convolutional-neural-networks (accessed Sep. 03, 2023).

[17] “Convolutional Neural Networks — A Beginner’s Guide | by Krut Patel | Towards Data Science.” https://towardsdatascience.com/convolution-neural-networks-a-beginners-guide-implementing-a-mnist-hand-written-digit-8aa60330d022 (accessed Sep. 04, 2023).

[18] “What is Image Segmentation: The Basics and Key Techniques | Mindy Support Outsourcing.” https://mindy-support.com/news-post/what-is-image-segmentation-the-basics-and-key-techniques/ (accessed Sep. 04, 2023).

[19] M. Ali, M. Jabreel, A. Valls, M. Baget, and M. A. Mohamed, “LezioSeg: Segmenting Eye Lesions in Fundus Images Utilizing Deep CNN with Hybrid Multi-Scale Attention Schemes”, doi: 10.2139/SSRN.4396788.

[20] P. Porwal *et al.*, “Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research,” *Data 2018, Vol. 3, Page 25*, vol. 3, no. 3, p. 25, Jul. 2018, doi: 10.3390/DATA3030025.

[21] E. Decencière *et al.*, “Feedback on a publicly distributed image database: The Messidor database,” *Image Analysis and Stereology*, vol. 33, no. 3, pp. 231–234, 2014, doi: 10.5566/IAS.1155.

[22] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” *IEEE Trans Pattern Anal Mach Intell*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[23] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 2017, Accessed: Sep. 03, 2023. [Online]. Available: https://arxiv.org/abs/1704.04861v1

[24] E. Munuera-Gifre *et al.*, “Analysis of the location of retinal lesions in central retinographies of patients with Type 2 diabetes,” *Acta Ophthalmol*, vol. 98, no. 1, pp. e13–e21, Feb. 2020, doi: 10.1111/AOS.14223.

[25] “Messidor - ADCIS.” https://www.adcis.net/en/third-party/messidor/ (accessed Sep. 03, 2023).

[26] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” *BMC Med Imaging*, vol. 15, no. 1, pp. 1–28, Aug. 2015, doi: 10.1186/S12880-015-0068-X/TABLES/5.

[27] “TensorFlow.” https://www.tensorflow.org/?hl=es-419 (accessed Sep. 03, 2023).

[28] “Keras: Deep Learning for humans.” https://keras.io/ (accessed Sep. 03, 2023).

[29] “OpenCV: OpenCV modules.” https://docs.opencv.org/4.x/ (accessed Sep. 03, 2023).

[30] “Pillow (PIL Fork) 10.0.0 documentation.” https://pillow.readthedocs.io/en/stable/ (accessed Sep. 03, 2023).

[31] “API Reference — scikit-learn 1.3.0 documentation.” https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics (accessed Sep. 03, 2023).

[32] “NumPy.” https://numpy.org/ (accessed Sep. 03, 2023).

[33] S. Vujosevic *et al.*, “Screening for diabetic retinopathy: new perspectives and challenges,” *Lancet Diabetes Endocrinol*, vol. 8, no. 4, pp. 337–347, Apr. 2020, doi: 10.1016/S2213-8587(19)30411-5.

8 Annex

8.1 Counting Lesions Functions

```
def find_OD(contours):
    best_contour = None
    best_circularity, best_per, best_area = 0, 0, 0
    for contour in contours:
        area = cv2.contourArea(contour)
        perimeter = cv2.arcLength(contour, True)
        if perimeter > 100: #We apply a high threshold to
delete some noise that could be generated (small predicted
circles)
            circularity = (4 * math.pi * area) / (perimeter *
perimeter)
            if circularity > best_circularity:
                best_circularity = circularity
                best_contour = contour
                #best_per, best_area = perimeter, area
    #print(best_per, best_area)
    return best_contour


def count_obj(img,img_od):
    #Find OD position
    contours, _ = cv2.findContours(img_od, 2, 1)
    od = find_OD(contours)
    center, radius = cv2.minEnclosingCircle(od)
    x_od, y_od = center

    #Count total number of lesions
    contours, _ = cv2.findContours(img, 2, 1)
    total = str(len(contours))
    y, x = round(img.shape[0]/2), round(img.shape[1]/2)

    #Count number of lesions per quadrant
    upperNasal = img[0:int(y), 0:int(x)]
    contours, _ = cv2.findContours(upperNasal, 2, 1)
    UpNas_objects = str(len(contours))

    lowerNasal = img[int(y):img.shape[0], 0:int(x)]
    contours, _ = cv2.findContours(lowerNasal, 2, 1)
    LNas_objects = str(len(contours))
```

```
        upperTemporal = img[0:int(y), int(x):img.shape[1]]
        contours, _ = cv2.findContours(upperTemporal, 2, 1)
        UpTem_objects = str(len(contours))

        lowerTemporal = img[int(y):img.shape[0],
    int(x):img.shape[1]]
        contours, _ = cv2.findContours(lowerTemporal, 2, 1)
        LTem_objects = str(len(contours))

        #If OD is on the right side, necessary to swap quadrants
    counts
        if x_od > round(img_od.shape[0]/2):
            UpNas_objects, LNas_objects, UpTem_objects,
    LTem_objects = UpTem_objects, LTem_objects, UpNas_objects,
    LNas_objects

    return UpNas_objects, LNas_objects, UpTem_objects,
LTem_objects, total
```

8.2 Grading Diabetic Retinopathy Function

```
def classify_DR(row, ma, hm):
    if row[ma] == 0 and ((row[hm] == 0) or (row[hm]== 1)):
        return 0
    elif  (0 < row[ma] <= 5) and row[hm] == 0:
        return 1
    elif (5 < row[ma] < 15) or (0 < row[hm] <= 5):
        return 2
    elif row[ma] >= 15 or row[hm] >= 5:
        return 3
    else:
        return -1
```

8.3 Get Confusion Matrix Function

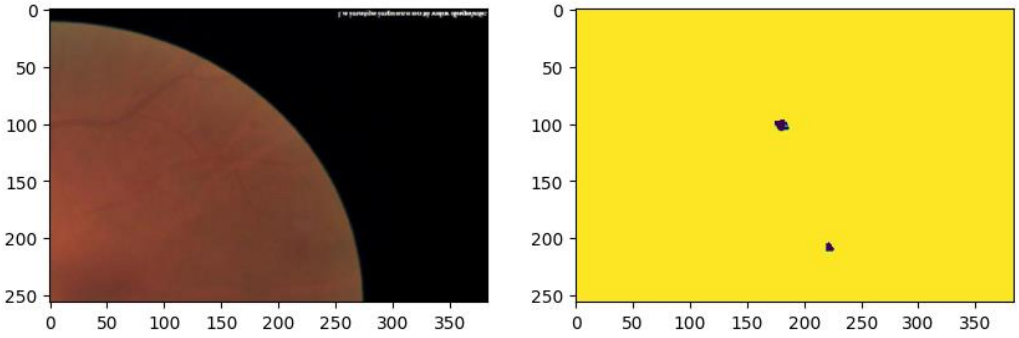
```
def obtain_CM(true, predicted, csv):
    df = pd.read_csv(csv)

    if (df[predicted].min() or df[true].min()) == -1:
        cm = np.zeros((5, 5), dtype=int)
        row = [true.split("_", 1)[1] + "_" + str(i) for i in
range(-1,4)]
        col = [predicted.split("_", 1)[1] + "_" + str(i) for i in
range(-1,4)]
```

```
        for true, pred in zip(df[true]+1, df[predicted]+1):
            cm[true, pred] += 1
        df = pd.DataFrame(cm, index=row, columns=col)
    else:
        cm = np.zeros((4, 4), dtype=int)
        row = [true.split("_", 1)[1] + "_" + str(i) for i in
range(4)]
        col = [predicted.split("_", 1)[1] + "_" + str(i) for i in
range(4)]
        for true, pred in zip(df[true], df[predicted]):
            cm[true, pred] += 1
        df = pd.DataFrame(cm, index=row, columns=col)
    return df
```

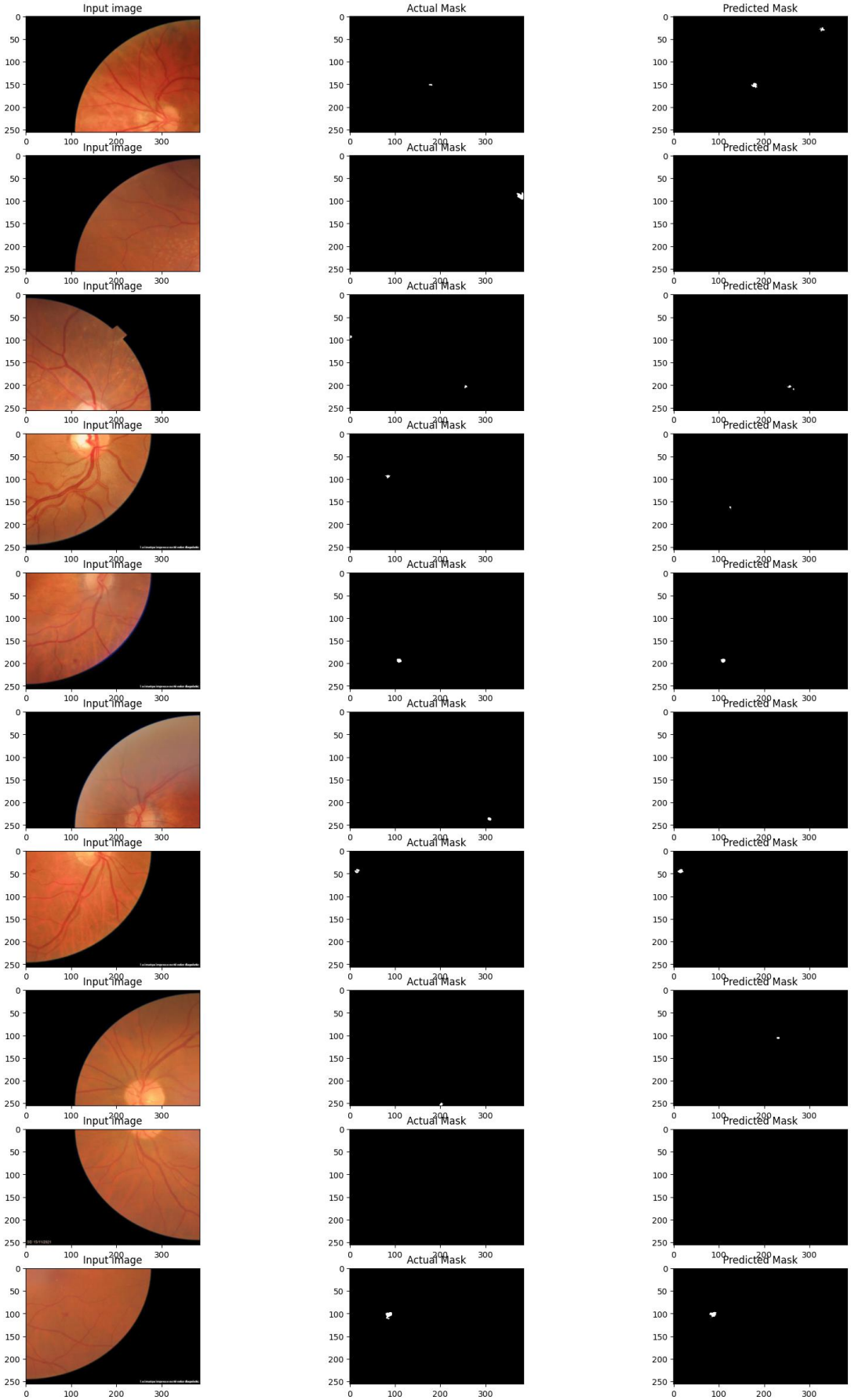
8.4 Training Dataset Configuration

```
resize = [256, 384]
tr_cfg = {
    'resize': resize,
    'hue_delta': 0.0,
    'scale': 1.0 / 255.0,
    'horizontal_flip': True,
    'vertical_flip': True,
    'rotate': True,
    'bright': True,
    'crop': False,
    'contrast': False,
    'gaus': False,
    'gray': False
     }
train_repeat = 12
```

8.5 Example of Image and Mask from the New Training Dataset



8.6 Example of the Validation Phase during Training



8.7 Example of the Test Phase