Idiap Research Institute

Master thesis
- Project report -

Computer-Aided Screening for Eye Diseases from 2-D Retinography

Author:
KHALIL Driss
Student number :
20-697-413
Submission date
June 19

Supervised by:
ANJOS André
TARSETTI Flavio

Year 2021 - 2022

Computer-Aided Screening for Eye Diseases from 2-D Retinography

Thesis performed in the

Biosignal Processing Group
Idiap Research Institute
programme of Artificial Intelligence
UniDistance

to obtain the degree of

Master of science in Artificial Intelligence

by

Driss KHALIL

Under the supervision of:

Dr. André Anjos, project supervisor
Flavio Tarsetti, company supervisor

Martigny, Idiap, 2023

Written by Driss KHALIL <driss.khalil@idiap.ch>

Dedication

"*In memory of my* ***grandmother*** *God bless her soul*"

Acknowledgements

First, I would like to thank Dr.André Anjos for his excellent supervision throughout this master thesis and for making it one of my most incredible learning experiences. I had the chance to meet Dr.André Anjos during the International Create Challenge, and I still remember all his advice throughout that challenge. Once I joined the AI master and saw that he was proposing a project, I immediately took the initiative to contact him to be my supervisor. I also want to thank my second supervisor Flavio Tarsetti, who was a great help to me in defining the project and motivation and also in writing the report. I still recall one of his remarks from our meeting that helped me avoid wasting a lot of time working on the wrong approach. I also would like to thank Olivier Bornet, who was here to help me during my master thesis, especially when I was still in Morocco waiting for the work permit. I also want to thank our deputy director François Foglia who was always here to support me with encouragement and advice. A very special thanks to all teachers and assistants at Idiap, who taught us the crucial elements we used in this thesis and answered all our questions as soon as they could. Finally, I would like to thank all the AI master students, in particular André Mayoraz, Sargam Shukla and Adi Niederberger, with whom I developed a great friendship.

Abstract

The human eye remains one of the essential organs in the human body; it can be affected by several diseases that can cause irreversible blindness. These diseases can appear without any visual symptoms; a late diagnosis would cause the loss of sight and cause blindness. This project's global goal is to detect those diseases using two-dimensional eye fundus images. Nevertheless, in all healthcare-related projects, interpretability is an essential element to care about. Detecting the eye elements is one of the ways to improve the interpretability, as the healthcare specialist will have an overview of how the model did classify the diseases.

In our work, we studied a new multi-task approach that could detect all the eye elements and classify the diseases using one model. As we wanted to explore more views on this approach, we focused in this project on segmenting two essential elements of the eye fundus, the vessels, and the optic disc, as they are the elements that are affected mainly by the diseases.

To address this, we disposed of several public datasets for each task. Firstly, we generated some baseline results using single-task approach models. Secondly, we analyzed the results provided by each model, and we decided to choose U-Net as the model we want to work within our Multi-task learning approach. All the datasets in our work were annotated on one specific task, making us look for a method that can work with disjoint datasets. The chosen method was based on alternating the training between the tasks at each epoch. We generated a ground truth for the non-annotated task using the predictor of the previous epoch. We tested the model on new datasets with different image quality for each experiment, allowing us to test our model's generalization. The chosen experiments were based on evaluating the model's performance using multiple images from different datasets and assessing the effect of having an imbalance of the images between the tasks. The metric used to compare the results was the F1-score.

From the baseline method, we remarked two problems: the enormous noise in the training and validation losses and the non-convergence compared to the single-task approach. Two methods were proposed to face those problems. The first one consisted of using gradient accumulation to update the model's weight once at each epoch. The second method was to generate the ground truth of the non-annotated task using the best epoch predictor.

The results have shown that the gradient accumulation approach decreased the noise from the losses. Still, the method based on the generation of the ground truth from the best epoch failed to help the model converge and caused even a higher loss than the baseline method.

Table of contents

List of Figures

List of Tables

List of Abbreviations

HCW Healthcare Workers

OCT Optical Coherence tomography

MTL Multi-task learning

AMD Age macular degeneration

DR Diabetic Retinopathy

OD Optic-disc

OC Optic-cup

AUC Area under the curve

TP True positive

TN True negative

FP False positive

FN False negative

CNN Convolution neural network

FCNN Fully Convolution neural network

HED Holistically-Nested Edge Detection

DRIU Deep Retinal Image Understanding

Chapter 1

Introduction

Artificial Intelligence (AI) has become a major frontier in healthcare and has proven important contributions in automatic diagnosis and treatment of diseases. AI has provided healthcare institutions and practitioners with tools that they can use to reduce pressure on their workloads and redefine their workflows, and let them focus on the tasks that needs more reflection. [78]. However, while some algorithms outperform healthcare workers (HCW) in a variety of tasks, they are not yet fully integrated into day-to-day medical practice due to several regulatory concerns as the European Commission has classified medical-AI as a high-risk field that needs a lot of attention before integrating it into real-life clinical practice [58].

AI approaches in healthcare today provide new ways to assist medical experts. Due to ever increasing digitization processes and healthcare costs, HCW spend more and more time analyzing and interacting with multiple (digital) patient data. To avoid the explosion of costs, reliance on new screening and diagnosis techniques based on statistical analysis seems preferable to increasing the ratio of HCW per patient, especially in low-income regions. In this context, Computer-Aided Diagnosis (CAD) are systems that can assist HCW in the analysis and interpretation of medical data supporting disease detection and grading. These systems have attracted extensive attention from clinicians and researchers alike, as they may provide valuable insights, also offering early screening for patients [86]. AI-based CAD has made very significant progress over the last decade due to the increase of computational power, and the high access to large volumes of data, in addition to the maturity of new AI algorithms.

In this thesis, we focus on the use of Machine Learning (ML), and specifically Deep Learning (DL) for the early detection of eye diseases which can silently cause irreversible blindness. Healthcare systems are struggling to give satisfactory eye care, especially on an aging population. As a natural side effect, this trend projects an increased level of blindness from major diseases in 2050 with a projection of 114,6 million blind people in an expected population of 9,69 billion [7]. Multiple eye diseases are asymptomatic, but display severe consequences [56]. Early detection can help in early treatment which may prevent sight loss.

Large scale eye-fundus screening remains one of the most important challenges for sight loss prevention [83]. The ability to image the retina (the light-sensitive tissue), and develop techniques for analyzing such images is, therefore, of great interest. There are two major techniques to do so:

- **Retinography:** is a 2-D representation of the 3-D retinal semi-transparent tissues projected onto the imaging plane, which is obtained using reflected light. This technique

(a) Table-top (10-20kCHF) (b) Portable (phone-attachable, 5kCHF)

Figure 1.1: Eye fundus cameras for 2DFI systems.

is the cheapest to manufacture and commercialize, and has received increased attention due to its widespread use, accuracy and availability in clinical settings.

- **Optical Coherence tomography (OCT):** is a non-invasive imaging test that uses light waves to take cross-section pictures of the patient's retina. OCT exams are typically used in follow-up care of eye conditions and precision diagnostics. These devices remain at least an order of magnitude more expensive than traditional retinography cameras, and possibly 20 times more expensive than a portable version see Figure 1.1a and 1.1b.

We dedicate this work to the analysis of retinographic (2-D) images, given the availability of public datasets for such task see Table 2.7, and its reduce costs for mass screening contexts, where automated analysis boosted by AI systems may provide its highest impact. Figure 1.2 shows a typical image of the eye fundus via Retinography, together most visible elements.

The eye remains one of the most complex organs in the human body. Vision is possible due to interaction of multiple components [80]. Multiple diseases can badly affect such elements and cause sight loss. In our work, we will focus our attention to three principal diseases that constitute one of the leading cause for blindness worldwide [19]: Age-related Macular Degeneration (AMD), Diabetic Retinopathies (DR) and Glaucoma.

Age-related Macular Degeneration (AMD) corresponds to the deterioration of the macula [3], which is the small central area of the retina of the eye that is responsible for our visual acuity (see Figure 1.2) with the appearance of multiple lesion there. If severe, it can result in loss of vision in the middle of the visual field. With time, the patient may develop a complete loss of central vision. It is possible to distinguish between two types of AMD which presents different symptoms as illustrated in Figure 1.3. The first type of AMD is **Dry Macular** that causes a formation of drusen and yellow pigmentation, and Geographic atrophy in the macula. The second one is **Wet Macular**, where abnormal blood vessels behind the retina start to grow under the macula [82].

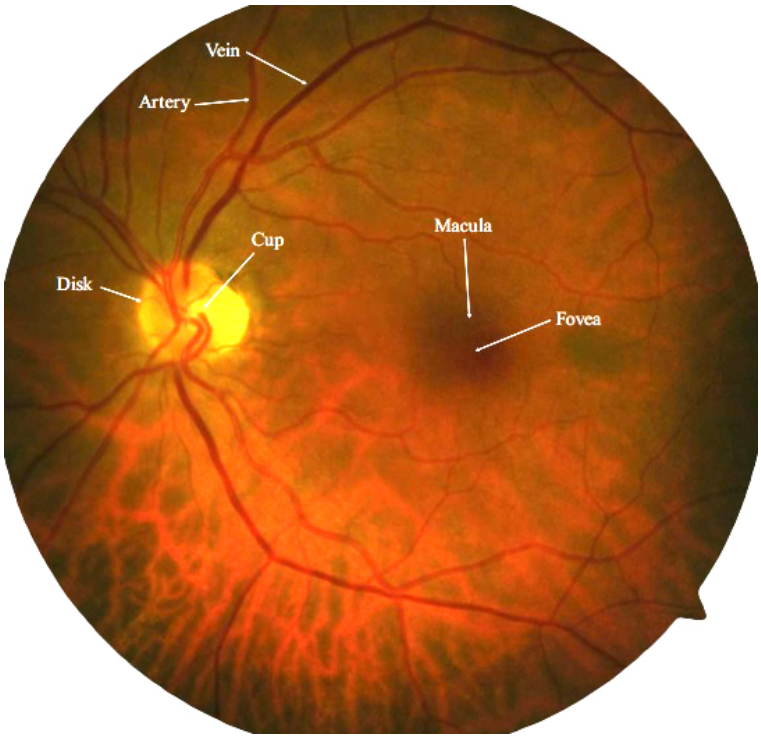


Figure 1.2: Most notable elements of the eye fundus.

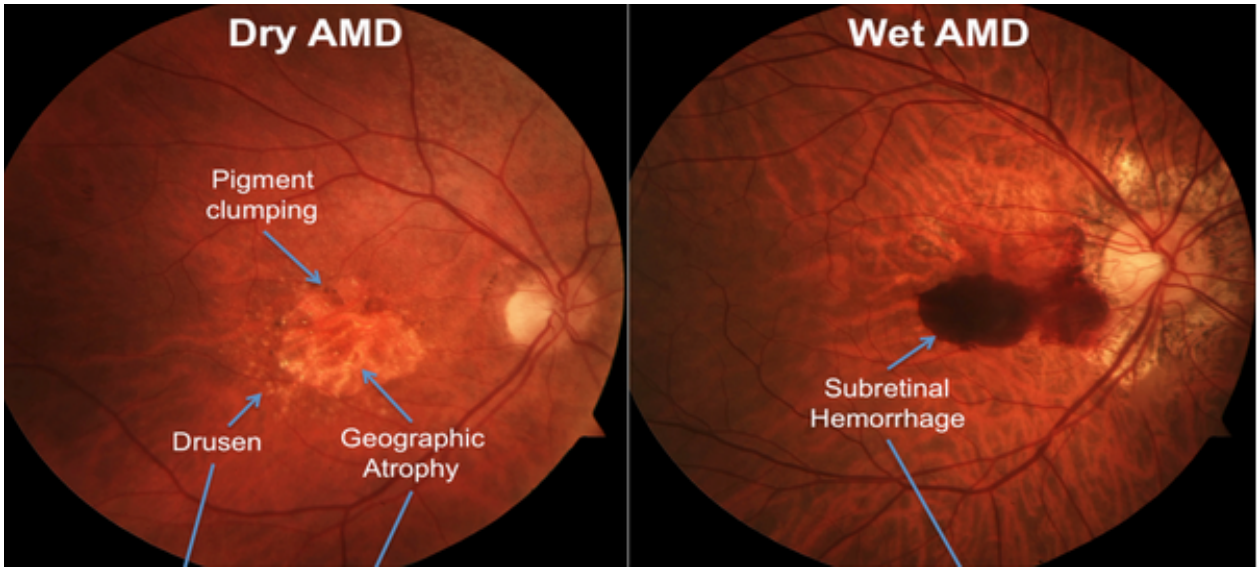


Figure 1.3: Eye fundus subject to dry and wet AMD and its most important signs.

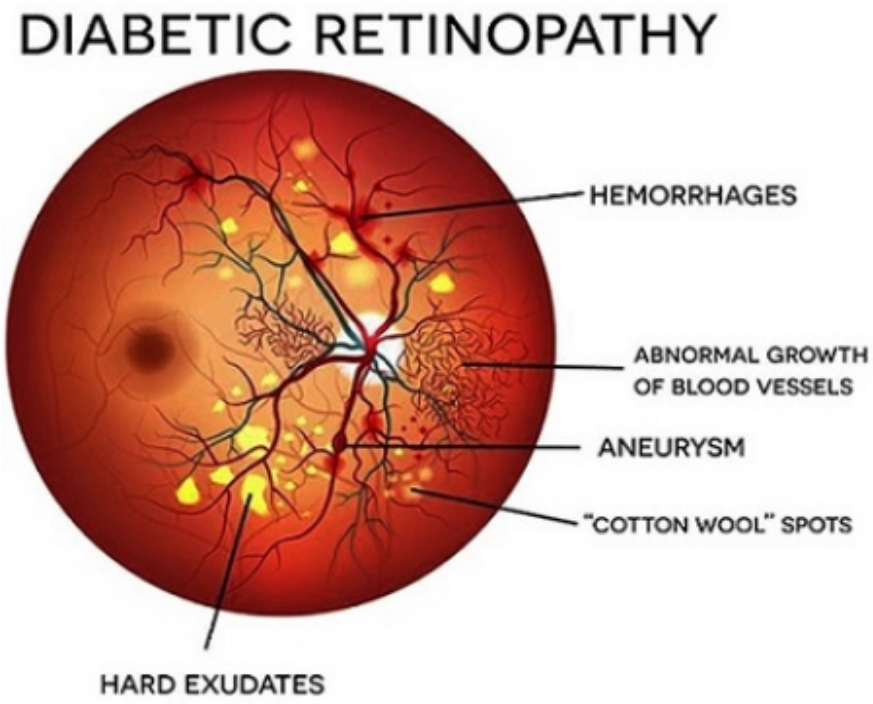


Figure 1.4: Illustration of common signs of diabetic retinopathies in the human eye.

Diabetic Retinopathies (DR) is a complication of diabetes, caused by high blood sugar levels damaging the retina [24]. DR detection consist of detecting several lesions in the eye of diabethic patients [18]. Figure 1.4 shows an illustration of a diabethic retinopathy eye diagram that is characterized by various abnormalities such as:

1. **Aneurysms**: small saccular dilations of capillaries that appears as round spots of dark red color with sharp edges in the retina background.
2. **Hard exudates:** deposits of lipids and proteins in the retina that produce bright lesions of yellowish-white color with prominent and irregular edges.
3. **Cotton wool spot:** a result of accumulations of axoplasmic material within the nerve fiber layer that appear as fluffy white patches on the retina.
4. **Hemorrhages:** wide accumulation of blood in the retina.
5. **Abnormal blood vessels** in the retina.

Glaucoma is a group of eye diseases which result in damage to the optic nerve and consequent vision loss. The most notable effect of Glaucoma is called cupping, in which the optical nerve is squeezed affecting vision acuity. Glaucoma is noticeable through the morphological analysis of the relationship between the optic disc (OD) and the optic nerve (ON) coupling, frequently referred to as the cup-to-disc ratio. Other measures are also used by ophthalmologists for diagnosis as neuroretinal rim loss, visual fields, and retinal nerve

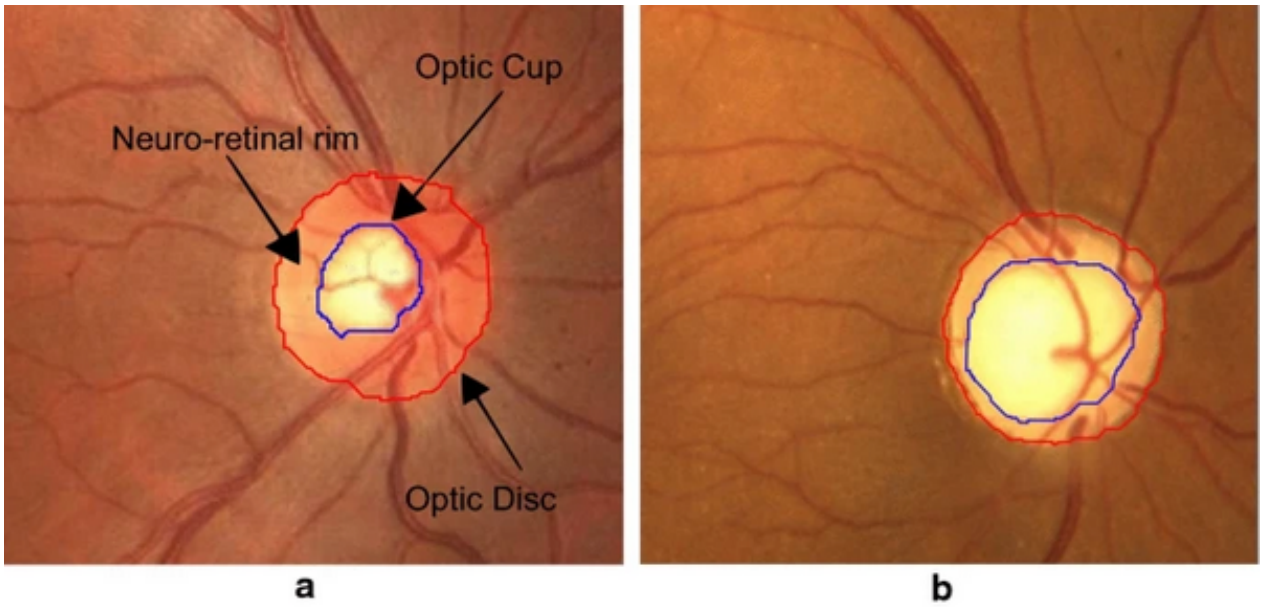


Figure 1.5: Digital fundus images cropped around the optic disc. **a-**Main structure of a healthy optic disc and **b-**Glaucomatous optic disc.

fiber layer defects [59]. Figure 1.5 shows the difference between the main structures of a healthy optic disc and a glaucomatous optic disc.

The ability to classify diseases with high accuracy remains one of the most popular areas in CAD. With the advent of DL, AI has been able to reach excellent levels of performance in various tasks, in a fraction of the time an expert would take to conduct a similar analysis [15].

There are two main approaches in literature for CAD from images (see Figure 1.6)

- **Direct** - a direct implementation, a CADx system is composed of a single block, in which the input image is directly fed to, outputting disease and grading without an intermediary segmentation and classification stage as shown in Figure 1.6.

- **Indirect** - the CADx system is composed of two building blocks as shown in Figure 1.6 (top). The input image extracted from the camera is input to a system responsible for the segmentation and classification of areas in the image where important structures are located. The segmented image is then fed into a classifier that is capable to relate the structural information to diseases, assigning a probability to each possible outcome.

While easier to implement, "direct" detection methods from images introduce interpretability concerns, which are rather difficult to surmount. On the other hand, "indirect" approaches in which radiological signs are first extracted from the image, are naturally more interpretable by HCW. However, implementing "indirect" approaches in practice may require prohibitive annotations costs, as experts need to precisely annotate large quantities of data to train DL models.

In our case, to deal with an indirect approach it is essential to do a segmentation of the retinal structure to detect the eye elements. The problem can be considered as a binary segmentation that requires precise annotation from experts (A time-consuming task). Nevertheless, the number of public datasets for such tasks has been increasing lately, although

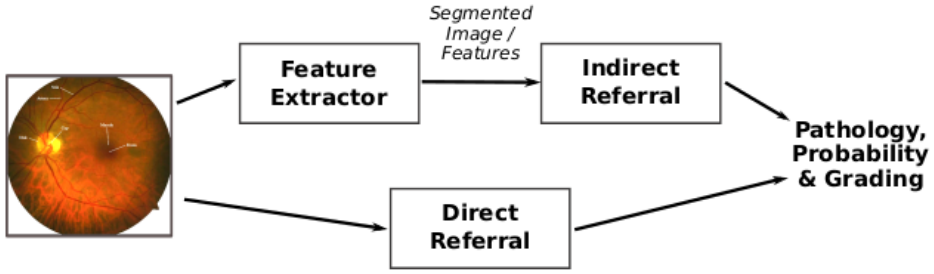


Figure 1.6: Schematics of an indirect (top) and direct (bottom) CADx systems operating on 2DFI images.

the number of images in some of them is still low.

In this thesis, we propose to gain in interpretability by exploring simultaneous delineation and diagnosis from a single image, via Multi-Task Learning (MTL). Image delineation will provide insights into anatomical elements, irregularities, and pathological findings. The diagnostic probability will contribute to the conclusion that can be obtained from such findings. We hypothesize that by leveraging on MTL, it will be possible to take advantage of the relatively small amount of annotated data for each task, to produce a strong MTL classifier for eye fundus screening.

Chapter 2

Related work

As already presented in Chapter 1, the objective of the global project is to detect three major diseases (Glaucoma, Diabetic retinopathy and AMD) which are the main reasons behind sight loss. A preliminary step to increase the performance and to help healthcare practitioners to interpret the results, is to apply a segmentation on 2DFI images such as Optic disc, Optic Cup segmentation, Lesion segmentation, and blood vessel segmentation before moving to the classification part. The following literature review will be structured as follows:

- **Segmentation**
  - Optic disc, Optic Cup segmentation
  - Lesion segmentation
  - Blood vessel segmentations
- **Classification**
  - Age Related Macular Degeneration (AMD)
  - Glaucoma
  - Diabetic retinopathy

At the end, we will conclude with an overview of existing multi-task learning (MTL) methods and show how we are going to approach a more interpretable CAD based on a MTL framework.

2.1 Optic disc (OD) and Optic Cup (OC) segmentation

Detecting the locations of the optic disc is a crucial task towards developing automatic diagnosis and screening tools for retinal diseases. The knowledge of the optic disc (OD) locations in the retina is considered essential for the diagnosis and screening of many retinal diseases [41].

There has been various studies conducted to determine the locations of the OD. Below, is a brief review of the major algorithms published in the literature for detecting the OD as illustrated in Table 2.1.

Lim et al [49] exploited a 9-layer CNN to accurately segment the optic disc and cup using feature exaggeration to extract relevant visual features. To quantify the the segmentation accuracy two metrics were used which are the non-overlap ratio (NOR) and the absolute vertical cup-to-disc (CDR). As the end-product was to build a screening model a Support Vector Machine (SVM) was applied on two features Cup to Disc Ratio and the predicted disc height and resulted on ROC curves. Overall the model based on CNN with Feature Exaggeration screening achieved a performance of (AUROC = 0.847).

In [54], the authors used a pre-trained Deep Retinal Image Understanding (DRIU) model inspired by the known GoogleNet architecture. They further applied data augmentation based on all the datasets using classical image transformations such as rotation and scaling. The authors tried to use these approaches to do both blood vessels and optic disc segmentation. The model was applied to each dataset separately and used the F1 score as a performance metric. The results obtained by the authors were as follows: 0.822,0.831 for blood vessel segmentation from DRIVE and STARE datasets respectively, and 0.971, 0.959 for optic disc segmentation on DRIONS-DB and RIM-ONE datasets respectively.

Feng et al. [27] presented a method based on data augmentation and Fully Convolutional Network (FCN) to detect accurate segmentations. The proposed model achieved end-to-end training, which consumes less time compared with patch-based training. The approach was applied to both Optic Disc and exudates segmentations and achieved the following results: - Sensitivity: 0.9312, 0.8135 - Specificity: 0.9956, 0.9876 - Positive Predictive Value: 0.8990, 0.8164 - F1-score: 0.9093, 0.8150 respectively in the optic disc and exudates segmentations.

Sevastopolsky at al [72] used U-Net CNN with Contrast Limited Adaptive Histogram Equalization (CLAHE) as a pre-processing step. The authors used two metrics: Intersection over Union (IOU) and F1 score for measuring the performance of their model and applied the model on 3 public datasets Drions-DB for optic disc segmentation with a performance of 0.94, and DRISHTI-GS for optic cup segmentation with a performance of 0.85, and RIM-ONE for both of them with a performance of 0.95 for OD and 0.82 for OC.

Another deep learning system based on FCN presented in [25]. The system predicted OD and cup boundaries in a single pass using fully convolutional networks and did not require resizing of the original image or cropping to OD region for cup prediction. The model was applied to both optic cup and optic disc segmentation and achieved a performance of 0.967 for optic disc and 0.897 for optic cup segmentation using F1-score as a metric.

Authors in [6] proposed a multi-scale sequential deep learning technique for the simultaneous detection of the center of the OD and fovea in color fundus images. The proposed method achieves an accuracy of 97%, 96.7% for the detection of the OD center and 96.6% and 95.6% for the detection of the foveal center of the MESSIDOR and Kaggle test sets respectively. Moreover, the method showed that the landmarks of an image can be localized in only 0.007s.

Another work presented in [57] trained CNN on full images to predict bounding boxes along with their analogous probabilities and confidence scores to detect OD. An amplification to enhance generalization capabilities on random samples was used as a pre-preprocessing step by adding multiple noises to enhance the robustness of the model. The authors applied their model to 4 public datasets MESSIDOR, EyePACS, DRIVE, STARE. DRIVE and STARE were only used in the testing part as they don't have annotations for optic disc segmentation. The method accomplishes an accuracy of 99.05% and 98.78% on the MESSIDOR and Kaggle datasets, and 99.41%, 98.37% for DRIVE and STARE datasets which were not used in the

Author	Method	Metric	Performance	Dataset
Lim et al. 2015 [49]	CNN Pixel-Level Classification; 9-layer	AUROC	0.847	MESSIDOR SEED-DB (Not available online)
Maninis et al. 2016 [54]	DRIU	F1 score	Blood Vessel : 0.822 Blood Vessel : 0.831 OD : 0.971 OD : 0.959	DRIVE STARE DRIONS-DB RIM-ONE
Feng et al. 2017 [27]	FCN with short and long skip connection	Sensitivity Specificity F1 score	93.12% 99.56% 0.9093	DRIONS-DB
Sevastopolsky 2017 [72]	U-net	F1 score	OD : 0.94 OD : 0.95 OC : 0.85 OC : 0.82	DRIONS-DB RIM-ONE DRISHTI-GS RIM-ONE
Edupuganti, Chawla, and Kale 2018 [25]	VGG16 FCN initialized weights from a model trained on ImageNet	F1 score	OD :0.967 OC : 0.897	Drishti-GS
Al-Bander et al. 2018 [6]	Multiscale sequential CNN + Contrast enhancement	Accuracy	OD : 97% ,96.7% Fovea : 96.6% ,95.6%	MESSIDOR EyePACS
Mitra et al. 2018 [57]	24-layered architecture of CNN	Accuracy	99.05% 98.78% 99.41% 98.37%	MESSIDOR EyePACS DRIVE STARE

Table 2.1: Optic disc, Optic cup segmentation state of the art results - AUROC :Area under the receiver operating characteristic

training phase.

2.2 Lesion segmentation

The detection of lesions facilitates in initial screening step, as they can represent the earliest signs of multiple eye diseases described in Chapter 1.

In the literature, many researchers were interested in detecting these lesions to add a layer of interpretability for eye disease diagnosis (Especially in DR). An overview of these researches is described in Table 2.2.

Shan et al. [73] aimed to detect the microaneurysm (MA) which is one of the lesions that help to detect DR using a method based on a Stacked Sparse Auto-Encoder (SSAE), the authors tried first to generate small patches from each image, then extract high-level features using the auto-encoder before feeding it to a softmax classifier. The dataset used for this approach was DIARETDB, and the resulting performances were as follow: Accuracy = 91.38%, AUC = 0.962 and F1 score = 91.34%.

Grinven et al. [33] presented a model to detect hemorrhages using a 9-layer CNN inspired by Oxfordnet [74] in addition to a selective sampling (SES) that aim to improve the speed and the performance of the model. The SES assigns a weight to each sample, the ones with the highest weight are selected in the next epoch. The weights are assigned using the classification error of the previous epochs. Two large datasets were used in this paper

with two different experiments, where the first one uses all the images of the datasets and the second one removes the poor quality images. The resulting performances for the first experiment were as follows: Sensitivity = 97.9%, Specificity = 91.4% and AUROC = 0.972 for Messidor dataset, and Sensitivity = 83.7%, Specificity = 85.1% and AUROC = 0.894 for EyePACS dataset.

Another approach was implemented by Orlando et al.[60] to detect red lesions on the eye using an Ensemble deep learning approach that extracts first some features using a CNN architecture and then augments them with some hand-crafted features which are fed to a random forest classifier. As the datasets used didn't include a mask for the fovea, the authors applied a threshold to the luminosity of the images to generate it. The authors presented two different types of analysis: DR screening (red lesions detection) and the need for referral to a specialist classification. The results for DR screening using AUC (ROC curve) are 0.9031, 0,8932 for e-OPHTA and MESSIDOR datasets respectively.

In their work Lam et al [47] applied a multi-class classification: (1) normal, (2) microaneurysms, (3) dot-blot hemorrhages, (4) exudates or cotton wool spots, and (5) high-risk lesions such as neovascularization, venous beading, scarring, and so forth using a CNN architecture inspired by GoogleNet with the help of a sliding window patch-based classification. A pixel-wise classification of the AUC (ROC curve) of 0.94 and 0.95, as well as a lesion-wise AUC (Precision/Recall curve) of 0.86 and 0.64, for microaneurysms and exudates, respectively were validated on e-OPHTA dataset.

Badar el al. [5] presented an end-to-end fully convolutional encoder-decoder-based network architecture for pixel-wise simultaneous classification of retinal pathology into object classes: 'exudates', 'hemorrhages', 'cotton wool spots', with the help of the patch-based learning approach. An accuracy of 99.24%, 97.86%, 88.65%, a sensitivity of 90.69%,80.93%,72.87%, a specificity of 99.64%, 98.54%, 89.32% for Exudates , HE , CWS respectively were achieved on Messidor dataset.

In the same way, Khojasteh et al.[45] proposed an approach for the identification of 'exudates', 'hemorrhages', 'microaneurysm' using a convolutional neural network architecture by embedding a preprocessing layer followed by the first convolutional layer to increase the performance of the convolutional neural network classifier in addition of the integration of two images enhancement technique CLAHE and contrast enhancement (CE). The model with Contrast enhancement had better results compared with the model with CLAHE. A sensitivity of 91%, 80%, 84%, and a specificity of 94%, 94%, 97%, and an accuracy of 91%, 82.5%, 78,8% were achieved on DIARETDB1 using the model with CE.

2.3 Blood vessel segmentation

The extraction of retinal blood vessels if done at early stages can be very helpful in diagnosing the severity of several eye diseases and accordingly the treatment can be followed [75].

Detecting vessels can be a really time consuming task done by healthcare workers. Many researchers have therefore proposed several methods to automate the segmentation of these blood vessels. Table 2.3 shows a summary of some research done previously in this field.

Maji et al. [53] proposed an Hybrid deep learning architecture to detect blood vessels from color fundus images. The method consist on a deep neural network (DNN) for unsupervised learning of vesselness dictionaries that use a sparse trained denoising auto-encoders (DAE).

Author	Method	Metric	Performance	Dataset
Shan and Li 2016 [73]	Stacked sparsed autoencoder (SSAE)	Accuracy AUROC F1 score	91.38 0.962 91.34	DIARETDB
Grinsven et al. 2016 [33]	9 layers CNN inspired by OxfordNet [74]	Sensitivity Specificity AUROC	91.9% , 83.7% 91.4% , 85.1% 0.972 , 0.894	MESSIDOR , EyePACS
Orlando et al. 2018[60]	CNN using LeNet + hand crafted features	AUROC)	0.9031 0.8932	e-OPHTA MESSIDOR
Lam et al. 2018 [47]	CNN using GoogleNet3	Feature AUROC AUC(PR curve)	MA, Exudates 0.95, 0.94 0.86, 0.64	EyePACS , e-OPHTA
Badar, Shahzad, and Fraz 2018 [5]	Encoder–decoder based FCN + Patch-based Learning	Feature Accuracy Sensitivity Specificity	Exudates , HE , CWS 99.24%,97.86%,88.65% 90.69%,80.93%,72.87% 99.64%,98.54%,89.32%	MESSIDOR
Khojasteh et al. 2018 [45]	CNN + CLAHE/Contrast enhancement	Feature Sensitivity CLAHE Specificity CLAHE Accuracy CLAHE Sensitivity CE Specificity CE Accuracy CE	Exudates , HE, MA 0.90 ,0.77, 0.82 0.94 , 0.94 , 0.94 90.2% , 83.3%, 67,8% 0.91 ,0.80, 0.84 0.94 , 0.94 , 0.97 91% , 82.5%, 78,8%	DIARETDB1

Table 2.2: Lesion segmentation state of the art results - hemorrhage (HE), microaneurysm (MA), Exudates (Ex) , Contrast enhancement (CE) , Contrast-limited adaptive histogram equalization (CLAHE)

The output is then fed to a random forest classifier that aim to detect vessels in the color image. The approach achieves the objective of vessel detection with max. avg. accuracy of 0.9327 and area under ROC curve of 0.9195 on DRIVE dataset.

Liskowski et al [51] implemented a deep neural network trained on a large (up to 400 000) sample of examples pre-processed with global contrast normalization, zero-phase whitening, and augmented using geometric transformations and gamma corrections, the authors have also tried to classify the blood vessels into two classes (thick and fine) vessels at the end of the paper. The author tried different architecture in his approach and achieved results up to 0.99 in AUROC and up to 0.97 in accuracy while testing on the three datasets (DRIVE, STARE, CHASE-DB).

In the same way, Lepetit-Aimon et al. [48] has proposed a novel Fully Convolutional Neural Network architecture capable of processing high-resolution images through a large receptive field at a minimal memory and computational cost. First, a single branch CNN is trained on whole images at low resolution to learn large-scale features. The architecture was used for both segmentation of blood vessels and the classification of artery/vein, and the results showed satisfying results for the segmentation part only. The model achieved an accuracy of 96.1% with a standard deviation of 0.3%. Adding Conditional Random Fields (CRF) decrase the performance on the segmentation part to 95.9%, but gives better results on the classification (Artery/Vein) part.

Authors in [14] proposed a novel two-stage vessel segmentation framework for vessel segmentation. A convolutional neural network (CNN) is used to correlate an image patch with a corresponding groundtruth which was reduced using Totally Random Trees Embedding (TRTE). Then the training patches are forward propagated through CNN to create a visual

Author	Method	Metric	Performance	Dataset
Maji et al. 2015 [53]	DNN + denoising auto-encoders (DAE) + RF	Accuracy AUC	93.27% 0.9195	DRIVE
Liskowski and Krawiec 2016 [51]	Preprocessing + CNN	Accuracy AUC	up to 97% up to 0.99	DRIVE STARE CHASE-DB
Lepetit-Aimon, Duval, and Cheriet 2018 [48]	Large Receptive Field Fully Convolutional Network	Accuracy	95.9%	MESSIDOR STARE DRIVE
Chudzik et al. 2018 [16]	CNN + generation of new outputs based on the structure of TRTE-based nearest neighbour search space	Sensitivity Specificity AUC	0.7881, 0.8269 0.9741, 0.9804 0.9646, 0.9837	DRIVE , STARE

Table 2.3: Blood vessel segmentation state of the art results

codebook. Finally, this codebook is used to builda generative nearest neighbour search space which define a number of neighbourhoods that regulate the process of generating annotations patches, often unseen during training process. The model achieves a sensitivity of 0.7881, 0.8269, and a Specificity of 0.9741, 0.9804, and an AUC (ROC) of 0.9646, 0.9837 for DRIVE and STARE datasets respectively.

2.4 Age-related macular generation (AMD)

Age-related macular generation (AMD) is one of the major causes of blindness, especially when diagnosed at a late stage. Being able to classify this disease from simple eye fundus images can be considered as one of the efficient solutions to prevent the blindness caused by this disease [38].

In the literature, various studies have been presented in this field, Table 2.4 presents an overview of the results of some of them.

Burlina et al. [11] studied the appropriateness of the transfer of image features computed from pre-trained deep neural networks to the problem of AMD detection. Features are extracted from a pre-trained deep convolutional neural network (DCNN) model used on the Imagenet dataset and then used as an input for a linear Support Vector Machine (LSVM) classifier to detect AMD. The authors used two types of datasets using AREDS with different number of images in each class of AMD (No AMD (1), Early AMD (2), intermediate AMD (3), advanced AMD (4)), and he used a binary classification by grouping the data as follows ({1&2} vs. {3&4}, {1&2} vs. {3}, {1} vs. {3}, {1} vs. {3 & 4}). The results showed good preliminary results (between nearly 92% and 95% accuracy) in both types of datasets in both groups.

In [10] Burlina et al. followed the same work by using a deep learning architecture based on Alexnet. The dataset was divided into two classes (No AMD + early AMD vs. intermediate and advanced AMD). To fit the expected AlexNet input size, a pre-processing phase was applied to all the images by detecting the outer boundaries of the retina, cropping images to the square that was inscribed within the retinal boundary and resizing the square to the AlexNET input size. The deep convolutional neural network method yielded accuracy that

Author	Method	Metric	Performance	Dataset
Burlina et al. 2016 [11]	Deep features with SVM	Accuracy	92 to 95%	AREDS
Burlina et al. 2017 [10]	DCNN using AlexNet	Accuracy AUC	91.6% 0.96	AREDS
Horta et al. 2017 [37]	DCNN + random forests	Accuracy Sensitivity Specificity AUC	79% 66.34% 88.95% 0.8476	AREDS
Govindaiah et al. 2018 [31]	A modified VGG16 neural network with batch normalization in the last fully connected layer	Accuracy	92.5%	AREDS
Grassmann et al. 2018 [32]	Ensemble network (Inception-ResNet-V2 Xception)	Accuracy	94.3%	AREDS KORA

Table 2.4: AMD classification state of the art results

ranged between 88.4% (Standard Deviation, 0.5%) and 91.6% (Standard Deviation, 0.1%), the AUC (ROC) was between 0.94 and 0.96.

Horta et al. [37] present a hybrid method based on random forests and deep image features to combine non-visual side-channel information with image data for classification. A preprocessing step was used for the images before applying the model, detection of the circular boundaries of the retina, and then cropping of the inscribed square of that circle was applied to the images. Then the normalization step, whereby the image is converted into the Lab color space("L" indicates lightness, "a" the red-green component, and "b" the yellow and blue components). The lightness channel (L) is processed to lower the intensity gradient of the image, in addition to a resizing of all images. The results showed an accuracy of 79%, a sensitivity of 66.34%, a specificity of 88.95% and an AUC (ROC) of 84.76% in the AREDS dataset.

Govindaiah et al. [31] presented a deep learning method to detect AMD. The method is based on a modified 16-layer neural network (VGG16), inspired by the winner of the ILSVRC-2014 competition, where batch Normalization is added in the last fully connected layers to improve generalization. The method experimented on both 4 class AMD and 2 class AMD. The authors achieved the best accuracy with the modified VGG16 network which reached 92.5% for the two-class problem with more than one hundred thousand images on the AREDS dataset. In the same problem Grassmann et al. [32] authors has proposed an ensemble learning technique with two deep neural networks, namely, Inception-ResNet-V2 and Xception with a custom fine-tuning approach to detect AMD. The authors defined 13 classes (9 AREDS steps, 3 late AMD stages, and 1 for ungradable images) and trained several neural network architectures on them. The approach used was divided into 4 steps: the first one is the pre-processing of images, the second one is to train multiple convolution neural nets (CNNs) independently. The third one consist of training a random forest algorithm to build a model ensemble based on the results of the single CNNs(step 2). The last step is to apply the model to the test data for AREDS and the KORA dataset. The algorithm detected 84.2% of all fundus images with definite signs of early or late AMD, and overall accuracy of 94.2% was achieved when counting all the classes.

Author	Method	Segmentation	Metric	Performance	Dataset
Chakravarty and Sivaswamy 2018 [14]	Multi-task CNN	OD segmentation by U-NET	AUC	0.9456	REFUGE
Perdomo et al. 2018 [63]	DCNN	OD, OC, using CNN	Accuracy AUC	89.4% using 0.82 using	RIM-ONE-v1 DRISHTI-GS1
Pal, Moorthy, and Shahina 2018 [62]	Deep Convolutional auto-encoder + CNN + multi-task learning	OD segmentation using U-NET	AUC	0.923	DRIONS-DB

Table 2.5: Glaucoma classification state of the art results

2.5 Glaucoma

Glaucoma is a disease that can be asymptomatic in its early stages and leads to a gradual but irreversible loss of sight. Early detection is therefore very important for sight loss prevention [23].

The literature showed a lot of interest for the classification of this disease, especially with the combination of Optic disc and cup segmentation as shown in Table 2.5.

Chakravarty et al. [14] presented a Multi-task Convolutional Neural Network (CNN) that jointly segments the Optic Disc (OD), Optic Cup (OC) and predicts the presence of glaucoma in color fundus images. The neural network used by the author utilizes a combination of image appearance features and structural features obtained from the segmentation. The results showed an F1 score of 0.92 for OD segmentation, 0.84 for OC segmentation, and an AUC of 0.95 on the task of glaucoma classification.

Authors in [63] present a multi-stage deep learning model for glaucoma diagnosis based on a curriculum learning strategy. This strategy aims to segment the optic disc and physiological cup, and then build a classification model with 3 classes ((healthy, suspicious, and glaucoma) using a deep convolution neural network and the segmentation result as an input, and achieved an accuracy of 89.4% using RIM-ONE dataset and an AUC 0.82 using DRISHTI-GS1.

Pal et al [62] proposed an architecture named G-EyeNet which is a deep learning multi-model network for glaucoma detection from retinal images. This architecture consists of a deep convolutional autoencoder combined with a CNN classifier which shares the encoder framework. This approach uses a multi-task learning procedure to minimize reconstruction and classification errors. The model achieved an AUC (ROC) of 0.923 on the DRIONS-DB dataset.

2.6 Diabetic Retinopathy

Diabetic Retinopathy is another eye disease that causes blindness. Diabetic patients have their retinas are examined regularly to check on the early signs of this disease, which can help them to avoid the late stage of it.

In the literature, multiple researches have proposed automatic solutions to detect this disease in an efficient and fast way. Table 2.6 shows an overview of some of the research done

in this field.

Abrámoffet et al. [2] implemented a system that uses a set of CNN-based detectors to each of the images. These detectors aim to detect normal anatomy, such as optic disc and fovea, as well as the lesions characteristic for DR, Though the CNNs are particular to each type of lesion and parameters vary slightly between them, they are inspired by Alexnet and the Oxford Visual Geometry Group. The results shows a sensitivity of 96.8% (95% Confidence Interval: 93.3%–98.8%), and a specificity of 87.0% (95% Confidence Interval: 84.2%–89.4%), and an AUC (ROC)of 98% (95% Confidence Interval: 96.8%–99.2%), with 6/874 false negatives, resulting in a negative predictive value of 99% on the Messidor-2 dataset.

Authors in [30] used a deep residual learning approach to detect DR which uses a customized deep convolutional neural networks for automated characterization of fundus photography. Using convolutional parameter layers, the network transforms images into features maps. Each of these layers propagates its output to the next layer. The model achieved a sensitivity of 94%, 93%, 90%, a specificity of 98%, 87%, 94% and an AUC (ROC) of 97%, 94%, 95% on the EyePACS, Messidor, e-OPHTA datasets respectively.

Another approach to detect DR was proposed by Quellec et al. [66], the authors built an interpretable method based on ConvNet by creating heatmaps that show which pixels in images play a role in the image-level predictions, the solution was also applied to lesion detection using DIARETDB datasets. For the DR detection part, results showed an AUC (ROC) of 0.954, 0.949 on the EyePACS and e-ophta datasets respectively.

Garcia el al [29] proposed a simple way to detect DR based on a VGG16 architecture already pre-trained from ImageNet. The approach is divided into two stages, the first one consists of the normalization of the saturation values of each figure, as well as the normalization of measurements and elimination of noise. The second one includes a parameter tuning approach for the CNN architecture. After training the results shows a specificity of 93.65% and an accuracy of 83.68% on validation process using EyePACS dataset.

In a different approach, Lin et al [50] tried to compare between using deep learning models directly on the original eye fundus image and using it on entropy images. The results showed that using entropy images can improve the deep learning performance and correctly detect the clinician-defined referable DR compared to normal images with an accuracy of 86.10% and a sensitivty of 73.24% and a specifity of 93.81% and an AUC (ROC) of 0.92.

2.7 Multi-task Learning

The proposed approach consists of building a Multi-task learning (MTL) framework leverage all information currently annotated for different tasks such as different kind of segmentation and classification. Due to the lack of annotation in multiple tasks in the data we will try to use the combination between several datasets partially annotated.

We will start by comparing our results with the approaches that are based on a single task, The approach will be tested first in an indirect way to improve the interpretability, if the performances are not satisfying as we wish, we will switch to an end-to-end approach.

Then we will try to compare our results with several models based on multi-task approach that aim to work on several task with similarities as we will present in the following parts.

In [52], authors proposed a multi-task deep neural network with spatial activation mech-

Author	Method	Metric	Performance	Dataset
Abràmoff et al. 2016 [2]	CNN AlexNet	Sensitivity Specificity	96.8% 87%	Messidor-2
Gargeya and Leng 2017 [30]	CNN deep residual learning	Sensitivity Specificity AUC	94%, 93%, 90% 98%,87%, 94% 0.97, 0.94, 0.95	EyePACS, MESSIDOR e-Optha2
Quellec et al. 2017 [66]	ConvNet	AUC	0.954 0.949	EyePACS, e-optha
García et al. 2017 [29]	CNN vgg16	Specificity Accuracy	93.65% 83.68%	EyePACS
Lin et al. 2018 [50]	CNN	Accuracy Sensitivity Specificity AUC	86.10% 73.24% 93.81% 0.92	EyePACS

Table 2.6: DR classification state of the art results

anism able to segment full retinal vessel (artery and vein simultaneously), without the pre-requirement of vessel segmentation. The performance were satisfying as they were equal to the state of the art for both task where the model achieves pixel-wise accuracy of 95.70% for vessel segmentation, and A/V classification accuracy of 94.50% for AV-DRIVE dataset, The algortithm has been tested also on INSPIRE-AVR dataset and achieved a skeletal A/V classification accuracy of 91.6%.

In [69], authors proposed a single model where given a retinal image it that can learn from separate datasets and predict all the classes. they have addressed this problem by utilizing multiple adversaries for addressing it as a multitask approach.

In [65], authors proposed a new approach based a convolutional multitask architecture with supervised learning and reinforcing it with weakly supervised learning. Three task are trained in the architecture: segmentation of red lesions and of bright lesions, and these two tasks will be done concurrently with lesion detection.

Zhao et al. in [87] proposed a novel multitask ensemble learning framework (DMTFs) to automatically perform optic disc quantification and achieve accurate multi-types multi-index OD quantification.

In [88], Zhao et al. proposed a weakly-supervised multi-task Learning method (WSMTL) to achive accurate evidence identification, optic disc segmentation and automated glaucoma diagnosis simultaneously.

In [79], Tu et al. proposed a multi-task network named feature Separation and Union Network (SUNet) for simultaneous DR (Diabetic retinopathy) and DME (diabetic macular edema) grading.

All the existing methods based on MTL still face several shortcomings. Performing different kind of segmentation and/or DR diagnosis/grading at the same time is a promising direction. The objective is to ensure a trade-off between the different tasks and more specifically Segmentation and classification. This is mainly because the high-level semantic features needed for the classification task tend to lack the spatial information required for segmentation.

2.8 Datasets

There is a growing number of datasets and public benchmarks available in current Fundus Images CADx literature. Table 2.7 provides, to the best of our knowledge, a close-to-exhaustive listing of publicly available datasets. The use of each dataset varies with respect to properties and diseases to be studied, but cover most of the needs for the 3 major eye conditions described earlier.

Datasets	Year	Annotations	Task
STARE [36]	2000	400 images, blood vessel annotation on 20 images	Optic disc, Optic Cup segmentation + Blood vessel segmentation
AREDS [17]	2003	Approx. 206,500 images	AMD
DRIVE [77]	2004	40 images, 33 normal and 7 mild DR	Optic disc, Optic Cup segmentation + Blood vessel segmentation
DIARETDBO [43]	2007	130 images, 20 normal and 110 DR	Optic disc, Optic Cup segmentation + Lesion segmentation
DIARETDB1 [44]	2007	88 images, 84 DR and 4 normal	Lesion segmentation + DR
DRIONS-DB [12]	2008	110 images, 23.1% chronic glaucoma and 76.9% eye hypertension	Optic disc, Optic Cup segmentation + Glaucoma
CHASE-DB [61]	2009	28 images	Blood vessel segmentation
e-OPHTA [22]	2013	47 images with exudates, 35 without. 233 normal images and 14 8MA images	Lesion segmentation
HRF [9]	2013	45 images 15 healthy , 15 DR , 15 glaucoma	Blood vessel segmentation
MESSIDOR [21]	2014	1200 images	Optic disc, Optic Cup segmentation + Lesion segmentation + Blood vessel segmentation + DR
Drishti-GS [76]	2014	101 images	Optic disc, Optic Cup segmentation + Glaucoma
AV-Paper-TMI [26]	2015	The paper extract Artety vein from 3 datasets AV-DRIVE : 40images AV-Inspire : 40 images AV-Wide : 30 images	Artery vein segmentations
KORA [8]	2016	Images from 2840 patients	AMD
IOSTAR [1] [85]	2016	30 images with a resolution of 1024 $\times$ 1024 pixels.	Vessel segmentation, Artery/Vein ratio, Optic disc segmentation
DRHAGIS [34]	2017	39 images of diabetic patient with 4 types of eye diseases, Glaucoma , Hypertension , DR AMD	Blood vessel segmentation
REFUGE [67]	2018	1200 annotated images	GLAUCOMA
EyePACS [42]	2019	35126 images	DR

Table 2.7: Public dataset for eye fundus

Chapter 3

Methods

This chapter describes the materials used to explore a Multi-task approach on two tasks with disjoint datasets. Exploring the literature, we found that most of the papers only use MTL with datasets annotated with all the tasks. In order to establish a baseline, we explore an MTL approach from the work presented in [35] which is based on knowledge distillation with a specific trick used to avoid forgetting. Using this method we generate baseline results using our chosen datasets. In the rest of the report, we will focus on two specific tasks which are optic-disc and vessel segmentation.

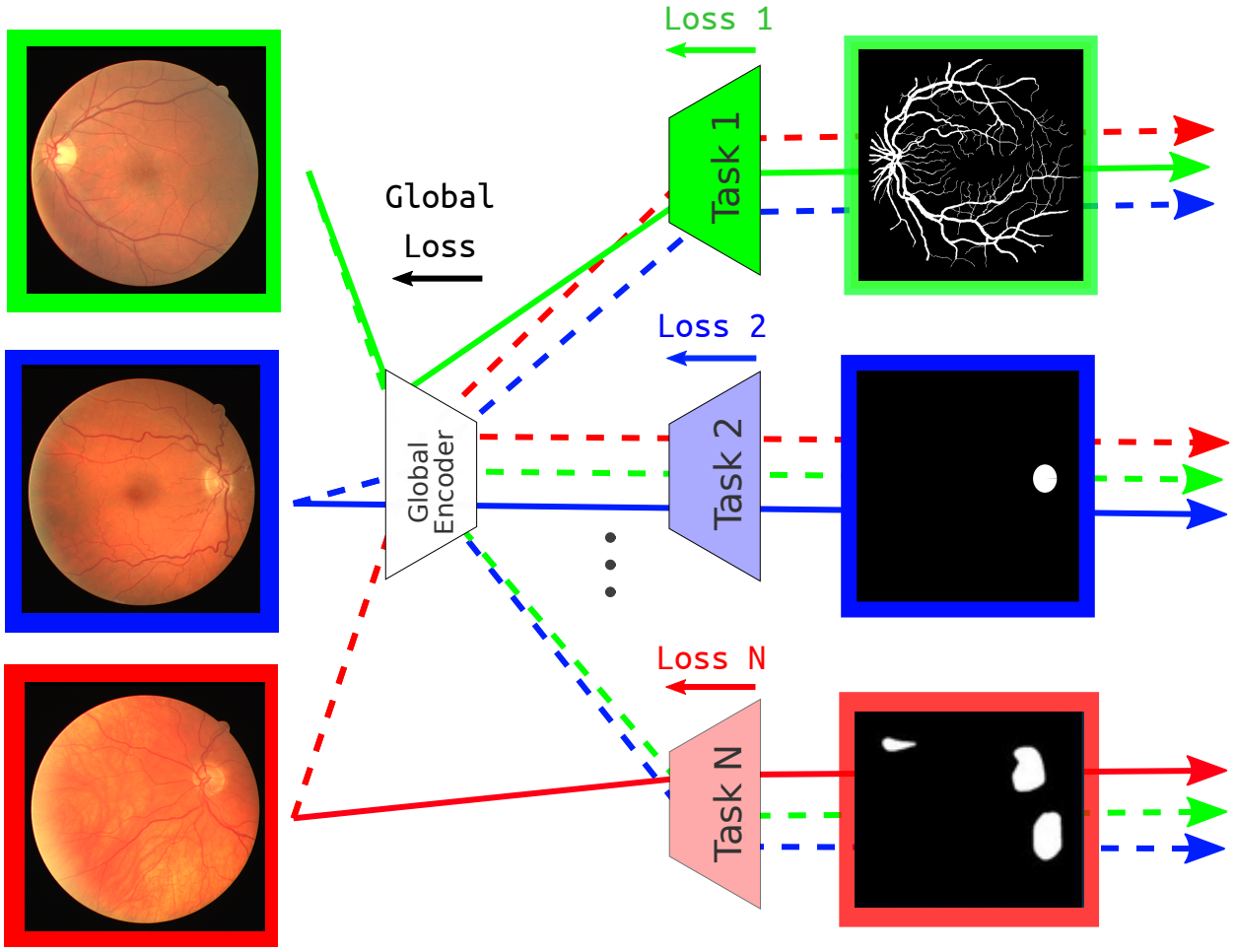


Figure 3.1: Global overview on the Multi-task method dealing with segmentation tasks of the eye fundus elements

3.1 Multi-task learning

In contrast to single-task learning, Multi-task learning enables the learning of several targets simultaneously, as we can see in figure 3.1. The idea behind multi-tasking learning came from the human learning process, as humans tend to learn single-tasks by integrating knowledge from several domains, which is the core of human intelligence. In analogy with deep learning, we can say that by sharing representation between related tasks, the model will aim to generalize and perform better than when training a single-task model [20].

However, the concept of learning a Multi-task model has never been easy and can introduce other problems that are not present in a single task learning [13]. This is because some tasks may have competing needs, which means that increasing the performance of one task will affect the performance of the other.

There are many different factors to consider when creating a shared architecture, such as the portion of the model parameters that will be shared between tasks, and how to parameterize and combine task-specific and shared modules.

A large portion of the MTL literature is devoted to the design of Multi-task neural network architectures and all the proposed methods have been partitioned into two groups

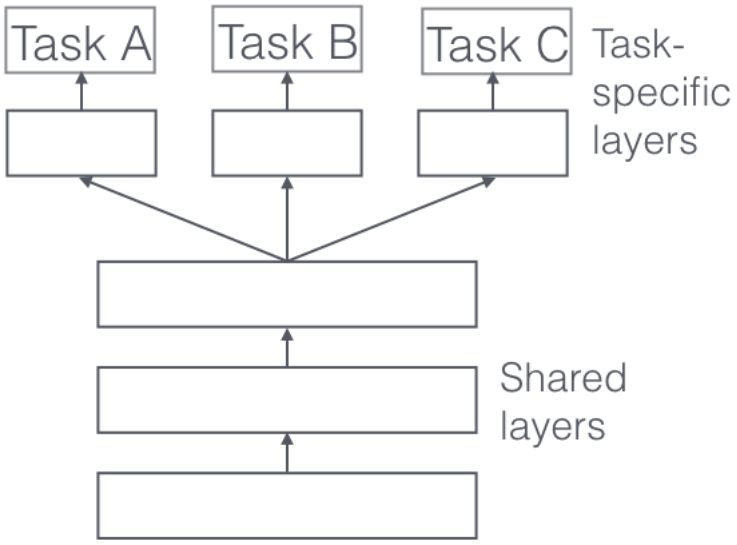


Figure 3.2: Hard parameter sharing for Multi-task learning in deep neural networks [4]

of Multi-task learning methods, hard parameter sharing and soft parameter sharing

3.1.1 Hard parameter sharing

Hard parameter sharing is the most commonly used approach in MTL. This method aims to share weights between all the tasks, which means that the weights are trained jointly to minimize multiple loss functions at the same time as depicted in figure 3.2. This shared feature space is used to model the different tasks with an additional specific layer learned independently for each task. This approach is known to have a low risk of overfitting since the tasks are learned simultaneously, which results in improved generalization. Nevertheless, this approach does not perform very well when we deal with non related close tasks.

3.1.2 Soft parameter sharing

In soft parameter sharing, the hidden layers and parameters have their own settings, but a regularization constraint that encourages similarities between the parameters is integrated to the model. In other words, it is like we have a model for each task but penalize the distance between the different model parameters. Although the model doesn't require any explicit parameter sharing, there is an incentive for the task-specific models to have similar parameters as illustrated in figure 3.3. This approach tends to give better performance, but is less scalable compared with hard parameter sharing.

3.2 Data

In this section, we describe the datasets used in our work, the targets and the pre-modelling operations.

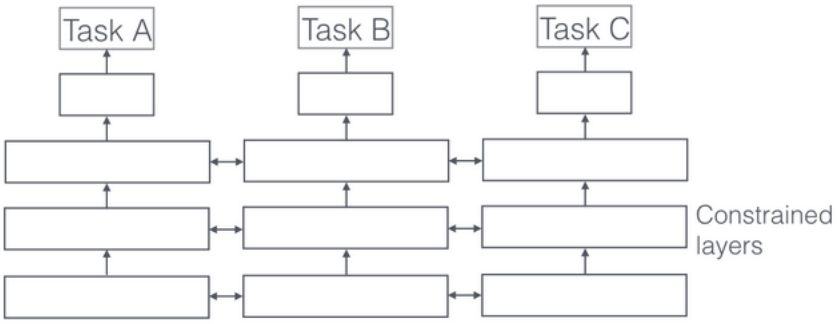


Figure 3.3: Soft parameter sharing for Multi-task learning in deep neural networks [4]

3.2.1 Target

In our work, we focus on the segmentation of the eye elements, and more specifically vessel and optic-disc segmentation, as they are two of the principal eye elements that are affected by the eye diseases as mentioned in chapter 1.

3.2.2 Datasets

As mentioned before, we focus on two specific tasks, which are optic-disc and vessel segmentation. For this purpose, we choose five datasets from each task. As the annotation for such tasks is really difficult, the number of images in each dataset is very low.

For vessels segmentations, we work with Drive, Stare, Iostar, Chasedb and HRF. Regarding the optic-disc task, we work with Iostar, drionsdb, refuge-disc, rimoner3, and drishtigs1. For more information about the datasets, see table 2.7 in chapter 2.

3.2.3 Data augmentation

It is common in machine learning problems that the more we have data, the more effective our model becomes, especially when we deal with adding lower quality images (adding noise) or images with random transformation. Our model becomes more robust against these different types of transformation [64].

In order to avoid overfitting and increase the robustness of our model, we use a data augmentation on all the training sets of the datasets. In our case, we only use two affine transformations, which are the rotation and the flipping.

3.2.4 Train-test split, normalisation and protocols

In this work, we create a common configuration for each dataset. The configuration consists on a split of 4 groups.

- Training set: it consists of the training images, as shown in table 2.7, each dataset having its proper training set.
- Testing set: it consists of the testing images, the set that will be used only in the testing phase.

- Training set with augmentation: it consists of the training images as well as the augmented images generated by the data augmentation technique.
- Validation set: it consists of the validation images used to train our model. In the case of the absence of a validation set in a specific dataset, we take the training set without augmentation as our validation set.

Each data point in the dataset is composed of an image, a ground truth, which is the label, and a mask. For the datasets without masks, we implement an algorithm based on transforming the image to binary with a threshold, in addition to using dilation and erosion techniques to generate the masks automatically to take only the eye part. However, for some datasets, such as drionsdb, the image was too dark to generate the mask with such an algorithm, so we tried to generate manually all the masks for all the images.

We prepare different protocols for the images before feeding the data to the models.

- Default protocol: in this protocol, we use padding and center cropping to transform the image into the nearest size dividable by 16.
- 512 resizing: we first pad the image into a square format before resizing all of the images to 512x512.
- 768 resizing: we first pad the image into a square format before resizing all of the images to a 768x768 shape.
- 1024 resizing: we first pad the image into a square format before resizing all of the images to 1024x1024 dimensions.

The second protocol is tested only for the optic-disc segmentation task, and the last protocol is tested only for the vessel segmentation task. Due to the large size of some datasets in optic-disc, we do not train the models on the default protocol.

After choosing the protocol to use, we then normalize all the images using a mean and standard deviation normalization for the images before feeding it to the models.

$$output_{channel} = \frac{input_{channel} - mean_{channel}}{std_{channel}}$$

3.3 Models

Before going deeper into the MTL approaches, we started first by using several segmentation models to get some baseline results for the single task approach.

3.3.1 DRIU

This model is based on CNN architecture that specializes in a base network for both vessels and optic disc segmentation from eye fundus images. This model is highly efficient at inference time as there is some specific layer depending on the task, as shown in figure 3.4.

As we can see, the architecture is composed of two parts, a base network and some specialized networks.

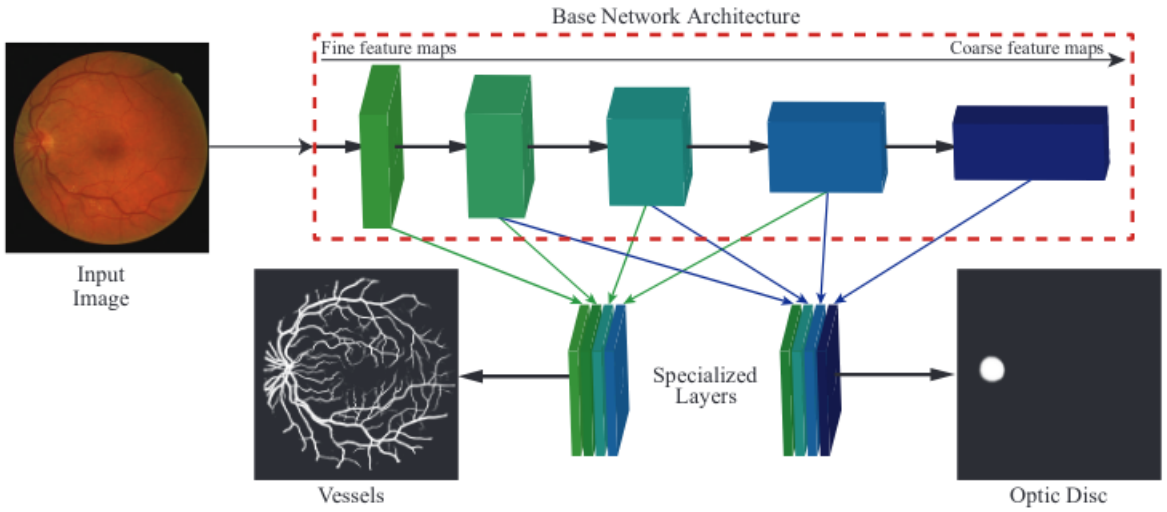


Figure 3.4: Overview of DRIU : Having a base convolution neural network architecture, the model extracts some side feature maps and explores specific specialized layers for vessels and optic-disc which are connected to a number of the features; the first four for vessels (left), and the last four for optic-disc (right). [55].

The base network is inspired by VGG, where they removed the fully connected layers from the end, and kept only the convolutional layers with RELU activations. As shown in figure 3.4 the architecture is based on four max pooling layers which separate the architecture into five stages. Each stage consists of several convolution layers. In order to keep the generalization of the architecture, the information becomes coarser as we move deeper into the network. All the layers of the base network are pre-trained on millions of images.

The end of the architecture consists of connecting the task specific layer (specific layer) to the final layer of each stage from the base network. The specialized layers produce feature maps in K different channels, which are then resized to the size of the original image, and finally there is a final convolutional layer which linearly combines all the feature maps from the volume created by the specialized layers into a regressed result.

As the architecture is designed for both vessels and optic-disc, there is a little difference concerning the choice of the layers from the base network for each of them. For vessel segmentation, the volume from the base network contains the four first stages, which are the finer ones, and for optic-disc the volume contains the last four stages, which are the coarser ones. This was done because when dealing with vessels, feature maps from the final stage erase the thin vessels, which can affect the performance of the model, and the finest ones don't help in detecting the discs.

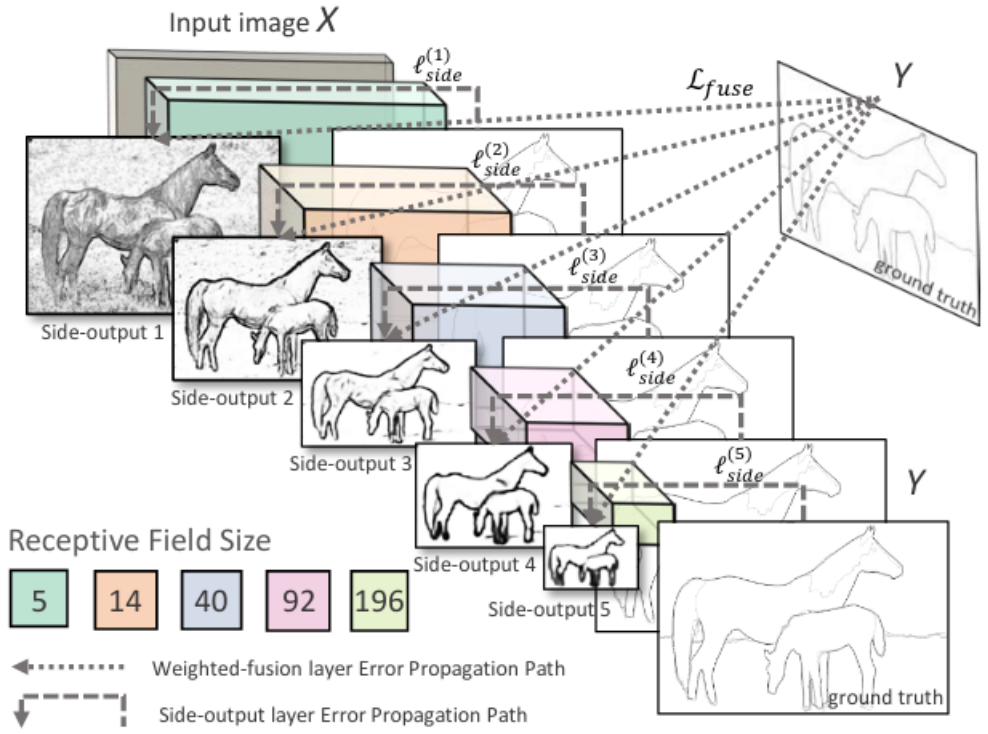


Figure 3.5: An illustration of the HED network architecture for edge detection, presenting the error backpropagation paths using an input image X. The outputs of HED are multi-scale and multi-level, where the side-output size decreases after each layer, and the receiptive field increases. A weighted fusion layer is added at the end to combine all the output from different scales. Deep supervision is forced after every side-output layer. The network is trained using different error propagation paths. [84].

3.3.2 HED

Holistically-nested edge detection (HED), is an architecture that aims to turn pixel-wise edge classification into image-to-image prediction by means of a deep learning model that leverages fully convolutional neural networks and deeply-supervised neural networks.

To create such an architecture, it was necessary to create a deep architecture to efficiently generate perceptually multi-level features and to have multiple stages with different strides, in order to capture the inherent scales of edge maps. Training such a network from scratch is a hard task. That is why the architecture has adopted VGGNet, which is trained on imagenet with the following changes to the architecture: they connected the side output layer with the last convolutional layer in each stage, and they cut the last stage of the VGGNet. The final HED architecture consists of five stages with a stride of 1, 2, 4, 8, and 16 with different receptive fields all nested in the VGGNet as shown in figure 3.5.

3.3.3 UNET

U-Net is a convolutional neural network that was developed for segmentation in biomedical imaging. The network is based on the fully convolutional network (FCN) and its architecture was modified and extended to work with fewer training images and to produce more precise segmentations. This architecture relies on a strong use of data augmentation which helps the model to use the available data more efficiently.

As we can see in figure 3.6, the architecture is composed of two stages, a contracting one and an expansive one. The contracting path consists of a normal convolutional neural network with repeated application of two 3x3 convolution operations with stride 2 for downsampling, and we double the number of feature channels in each downsampling. In the expensive path, there is an upsampling of the feature maps followed by a 2x2 up convolution that takes half of the number of the feature channels. Then they concatenate these features with the cropped features in the contracting path and two 3X3 convolutions followed by RELU. Finally, there is a 1x1 convolution layer that map each 64 component feature vector with the number of class.

3.3.4 M2UNET

The M2UNET is a new encoder-decoder inspired by UNET which is described above, but with an addition of pre-trained components from MobileNetV2 in the encoder and a new contractive bottleneck block in the decoder part. MobileNetV2 is an architecture based on an inverted residual structure where the input and output are thin bottlenecks, which is the opposite of traditional residual models. As we can see, M2unet is a combination of UNET and MobileNetV2 which takes the strengths of UNET and the speed of MobileNetV2.

M2UNET is divided into two parts. The first one consists of an encoder that takes the first 14 layers of MobileNetV2 pre-trained on ImageNet. This part is composed of two principal blocks, which are shown in figure 3.7 and introduced in [71]. These blocks are inverted bottlenecks, where the one with stride = 1 contains a residual connection if only the input and output have the same number of channels, the other bottleneck with stride = 2 doesn't have any residual connection. Both of them start with a convolution 1X1 that expands the number of the channel by a factor of t, followed by a depthwise separable convolution based on a depthwise and a pointwise convolution.

The second part is the decoder, which consists of using the same bottleneck with stride = 1 but with a contracting factor, without any residual connection, and instead of using transposed convolutional operations, they used a bilinear upsampling to upsample the spatial resolution of feature maps.

We can see in Figure 3.8 an example of how the model performs on an image with a shape of 544x544x3 from the encoder to the decoder part.

3.3.5 LWNET

This architecture is based on a small number of parameters (78k) compared with the previous one and it closely matches their performance. The idea behind this architecture is to use two modified UNET models with around 34k parameters each to create a strong network that outperforms the complex models. The input image is forward passed in a standard small

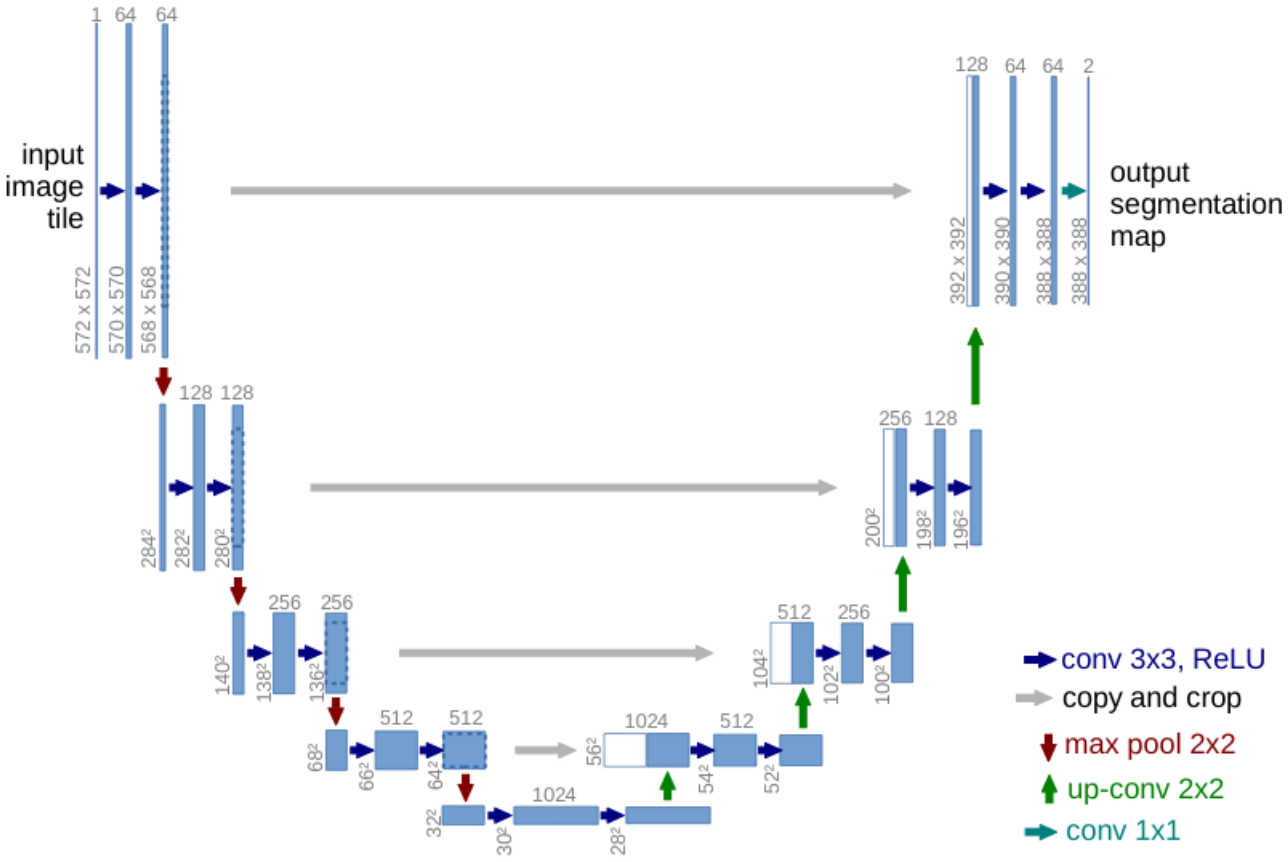


Figure 3.6: An Overview of the U-net architecture on a 32x32 image Each blue box corresponds to a multi-channel feature map. The width of the box represents the number of channels, and the height of the box represents the image resolution at that specific time. White boxes represent copied feature maps. The arrows denote the different operations. [68].

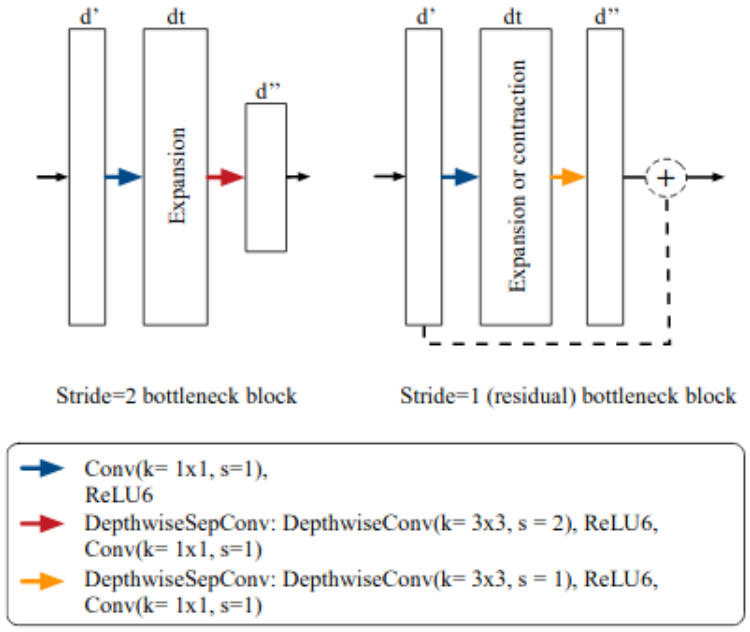


Figure 3.7: Bottleneck blocks: each box represents a multi-channel feature map. The height corresponds to the resolution, the width to the number of channels [46].

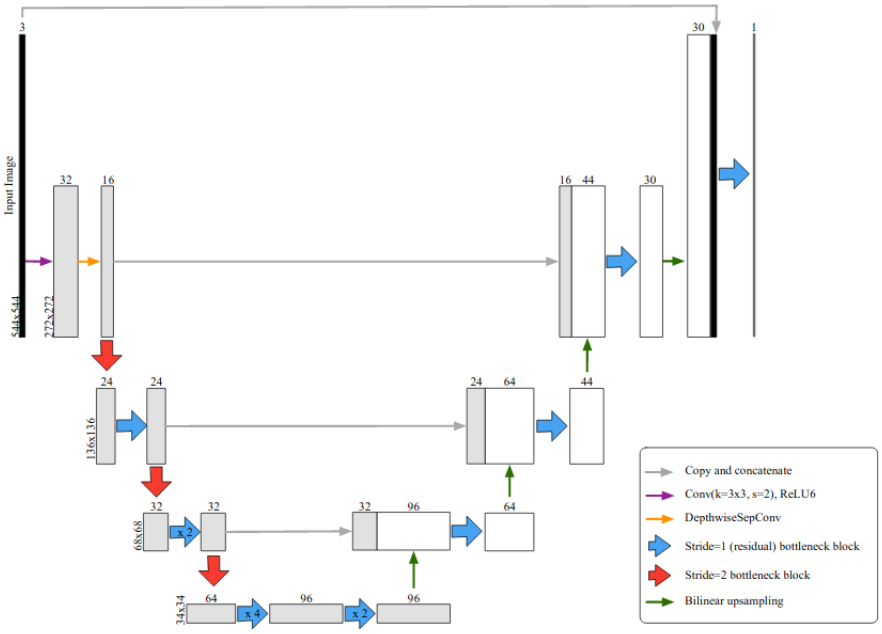


Figure 3.8: M2unet architecture on a 3x544x544 image (from the drive dataset). Each white and grey rectangular box represents a multi-channel feature map. The height of each block represents the resolution of the image at that time, and the width represents the number of channels. The feature maps in the encoder part are represented in grey [46].

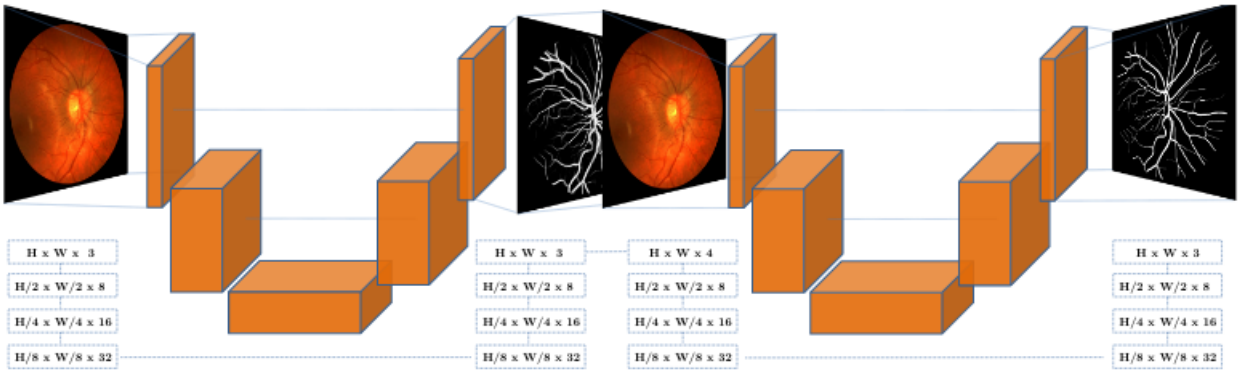


Figure 3.9: Representation of the WNet architecture on an eye fundus image The network is composed of 2 U-nets with around 70k parameters and achieves comparable results with complex architecture [28].

unet and the result is concatenated with the input image and then repassed in through the second unet as represented in figure 3.9. This architecture aims to create a first output for the model via the left part of the LWNET and then create a focus on the most important areas in the image so that it generates better results in the second pass.

3.4 Losses

3.4.1 SoftJaccard Binary Cross Entropy Logits Loss

This loss [40] is a combination of two well-known losses, the binary cross entropy and the softJaccard loss. The loss used here was inspired by [39] where they proposed a way of putting the Jaccard index in a differentiable form which facilitates the optimisation of the loss during the training.

The Jaccard index can be represented with the following formula:

$$J(A,B) = \frac{A \bigcap B}{A \bigcup B} = \frac{A \bigcap B}{|A| + |B| - |A \bigcap B|}$$

which can be represented with the following formula when dealing with images.

$$J = \frac{1}{n} \sum_{c=1}^{2} \omega_c \sum_{i=1}^{n} \frac{y_i^c \hat{y_i^c}}{y_i^c + \hat{y_i^c} - y_i^c \hat{y_i^c}}$$

where y_i^c, $\hat{y_i^c}$ are the binary value (label) and the predicted probability for the pixel, and ω_c will be equal to 1 to facilitate the calculation.

The combination of this Jaccard index with the Binary Cross Entropy gives the following formula.

$$L = \alpha H + (1 - \alpha) * (1 - J)$$

where H is the binary cross entropy loss , and α was chosen to be equal to 0.7 while using this loss.

3.4.2 Weighted Binary Cross Entropy Logits Loss

This loss was presented in [55] and was tested in the Driu article. It has the advantage when using unbalanced classes. Like in our case, when dealing with vessels or optic-disc segmentation, only 10% of the ground truths represent what we want to detect, and the other part is the background. This loss is based on [81] which was originally done for tasks such as contour detection from natural images. Considering X the input and $Y_n = y_j^n, j = 1, ..N$ with $y_j^n \in \{0, 1\}$ the predicted pixel-wise labels. The final loss function is then defined as :

$$L = -\beta \sum_{j \in Y_+} log(P(y_j = 1|X) - (1 - \beta) \sum_{j \in Y_-} log(P(y_j = 0|X)$$

The multiplier $\beta = \frac{|Y_-|}{|Y_+|}$ is the key element to handle the imbalance of the classes. The probability P(.) is obtained using sigmoid activation on the final convolution layer.

3.5 Metrics

3.5.1 Precision-Recall Area Under Curve (PR AUC) Score

We start first by explaining the precision and recall metrics before presenting the AUC metric.

Precision quantifies the number of correct positives made. In other words, it is the number of true positives over the number of true positives plus false positives.

$$Precision = \frac{Tp}{Tp + Fp}$$

Recall quantifies the number of correct positive predictions made out of all the positive predictions that could have been made. In other words, it is defined as the number of true positives (Tp) over the number of true positives plus the number of false negatives (Fn).

$$Recall = \frac{Tp}{Tp + Fn}$$

This metric is more useful as it is only concerned with the positives, which is the minority class, and in such a problem we are more interested in detecting the elements of the eye rather than detecting the background [70].

The PR curve combines precision and recall in a single plot. For every threshold, you calculate the two metrics and plot them. The higher on the y-axis your curve is, the better your model performance. In order to get a number to describe the model performance, we can use AUC, which is the area under the curve. We can think of PR AUC as an average precision score calculated for each recall threshold.

3.5.2 F1-score

This metric is more useful when dealing with imbalanced classes as it helps to balance the metric across both positive and negative pixels. F1-score is a metric that combines both precision and recall, but in a way that the metric will give a good performance only if both the recall and precision are high. The F1-score is also presented as the harmonic mean of precision and recall,

The F1-score can be calculated following this formula:

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

3.6 MTL approach for baselines

In this section, we present the single task baseline results using different protocols, and we also present a baseline approach that we followed during this project and some results following some experiences that we have chosen.

3.6.1 Single task baselines

Before trying models using Multi-task learning, it was mandatory to generate results for a single task approach with different protocols. As presented before, we worked with two different losses. In all the experiments that we will present, we used SoftJaccard BCE Logits Loss when using DRIU, UNET, and M2UNET, and we used the Weighted Binary Cross Entropy Logits Loss for LWNET. For each task, we tried three protocols for all the datasets, and we chose the best model for the tasks and generated a cross database results. The cross database results represent testing a model that was trained on a specific dataset on all the other datasets, which can help to see if the model is generalizable or not. Let's now present the results for both of the tasks. All the scores on the tables represent F1-scores with a threshold that was generated from the validation set.

Vessels segmentation

For this task, we used the default protocol and the protocols with resizing of 768 and 1024.

Table 3.1 represents the baseline results for vessel segmentation with the default protocol. As the datasets have different sizes, we tried here to use different batch sizes depending on the shape of the images.

Table 3.2 and table 3.3 represents the baseline results for the same task following the protocols with resizing 768 and 1024. As the datasets had the same size we used a common batch size for each model on all the datasets. The tables 3.4, 3.5, 3.6 presents the results of the cross database results for the 3 protocols.

After analyzing the results of each protocol, we tried to choose one model that gave correct results and another to see if the model provided a good generalization over all the datasets. Our choice was to use DRIU on the default protocol and UNET on the two others.

We can remark that having the datasets at the same size has improved the performance and the generalization of the models.

Datasets / Models	Driu	HED	M2unet	UNET	LWNET
Drive	0.821	0.813	0.802	0.825	0.828
Stare	0.828	0.815	0.818	0.828	0.839
Chasedb1	0.812	0.806	0.798	0.807	0.820
Hrf (1168x1648)	0.808	0.803	0.796	0.811	0.814
Hrf (2336x3296)	0.722	0.703	0.713	0.756	0.744
Iostar vessels	0.825	0.827	0.820	0.818	0.832

Table 3.1: Baseline results for vessels segmentation with the default protocol

Datasets / Models	Driu	HED	M2unet	UNET	LWNET
Drive	0.812	0.806	0.800	0.814	0.807
Stare	0.819	0.812	0.793	0.829	0.817
Chasedb1	0.809	0.790	0.793	0.803	0.797
Hrf	0.799	0.774	0.773	0.804	0.800
Iostar vessels	0.825	0.818	0.813	0.820	0.820

Table 3.2: Baseline results for vessels segmentation with the resizing 768 protocol

Datasets / Models	Driu	HED	M2unet	UNET	LWNET
Drive	0.813	0.806	0.804	0.815	0.809
Stare	0.821	0.812	0.816	0.820	0.814
Chasedb1	0.806	0.806	0.790	0.806	0.793
Hrf	0.805	0.793	0.786	0.807	0.805
Iostar vessels	0.829	0.825	0.817	0.825	0.824

Table 3.3: Baseline results for vessels segmentation with the resizing 1024 protocol

	Drive	Stare	Chasedb1	Hrf	Iostar vessels
Drive	**0.819**	0.759	0.321	0.711	0.493
Stare	0.733	**0.824**	0.491	0.773	0.469
Chasedb1	0.730	0.730	**0.811**	0.779	0.774
Hrf	0.702	0.641	0.600	**0.802**	0.546
Iostar vessels	0.758	0.724	0.777	0.727	**0.826**

Table 3.4: Cross database using default protocol on vessels dataset and DRIU as a model

	Drive	Stare	Chasedb1	Hrf	Iostar vessels
Drive	**0.814**	0.807	0.752	0.736	0.739
Stare	0.767	**0.829**	0.752	0.755	0.739
Chasedb1	0.774	0.800	**0.803**	0.730	0.771
Hrf	0.712	0.769	0.648	**0.804**	0.700
Iostar vessels	0.768	0.783	0.773	0.744	**0.820**

Table 3.5: Cross database using the resizing 768 protocol on vessels dataset and UNET as a model

	Drive	Stare	Chasedb1	Hrf	Iostar vessels
Drive	**0.815**	0.812	0.732	0.742	0.780
Stare	0.772	**0.824**	0.767	0.758	0.766
Chasedb1	0.774	0.762	**0.806**	0.729	0.779
Hrf	0.710	0.762	0.569	**0.807**	0.672
Iostar vessels	0.764	0.775	0.776	0.741	**0.825**

Table 3.6: Cross database using the resizing 1024 protocol on vessels dataset and UNET as a model

Optic-disc segmentation

For this task we used the protocols with the resizing 768 and 524. This choice was based on the analysis done [55], as we tried to use smaller image when dealing with optic-disc, and larger ones when dealing with optic-disc.

We can see from table 3.7 3.8 that the performance is a bit better when using smaller images, and we can see that the results for all the datasets present good performance. However, when we see the results in the cross-database approach as presented in tables 3.9 3.10 we can see that there is a problem when training the model on iostar-disc or testing on it. After some analysis, we concluded that this was due to the difference in the color of the optic disc in the different datasets, which is clear when looking at figure 3.10 and figure 3.11. All

Datasets / Models	Driu	HED	M2unet	UNET	LWNET	Driu OD
Drionsdb	0.958	0.961	0.960	0.961	0.922	0.960
Drishtigs1-disc	0.973	0.975	0.974	0.975	0.965	0.972
Iostar-disc	0.894	0.922	0.913	0.921	0.893	0.921
Refuge-disc	0.921	0.939	0.942	0.945	0.894	0.941
Rimoner3-disc	0.950	0.955	0.953	0.956	0.939	0.954

Table 3.7: Baseline results for optic disc segmentation with the resizing 512 protocol

Datasets / Models	Driu	HED	M2unet	UNET	LWNET	Driu OD
Drionsdb	0.945	0.917	0.959	0.960	0.875	0.949
Drishtigs1-disc	0.971	0.975	0.975	0.976	0.959	0.970
Iostar-disc	0.908	0.922	0.917	0.920	0.898	0.911
Refuge-disc	0.921	0.924	0.936	0.938	0.837	0.929
Rimoner3-disc	0.950	0.954	0.955	0.956	0.925	0.954

Table 3.8: Baseline results for optic disc segmentation with the resizing 768 protocol

	Drionsdb	Drishtigs1	Iostar-disc	Refuge-disc	Rimoner3-disc
Drionsdb	**0.961**	0.966	0.320	0.891	0.864
Drishtigs1-disc	0.953	**0.975**	0.057	0.923	0.884
Iostar-disc	0.050	0.086	**0.921**	0.074	0.115
Refuge-disc	0.885	0.943	0.057	**0.945**	0.879
Rimoner3-disc	0.864	0.890	0.064	0.734	**0.956**

Table 3.9: Cross database for optic disc segmentation with the resizing 512 protocol

the datasets have images with a light disc except iostar, where the disc is a bit green, so the model was either giving all the images as positive or all the images as negative.

	drionsdb	Drishtigs1	Iostar-disc	Refuge-disc	Rimoner3-disc
Drionsdb	**0.960**	0.957	0.057	0.886	0.888
Drishtigs1-disc	0.932	**0.976**	0.057	0.896	0.860
Dostar-disc	0.127	0.342	**0.920**	0.143	0.265
Refuge-disc	0.817	0.937	0.057	**0.938**	0.838
Rimoner3-disc	0.754	0.772	0.057	0.481	**0.956**

Table 3.10: Cross database for optic disc segmentation with the resizing 768 protocol

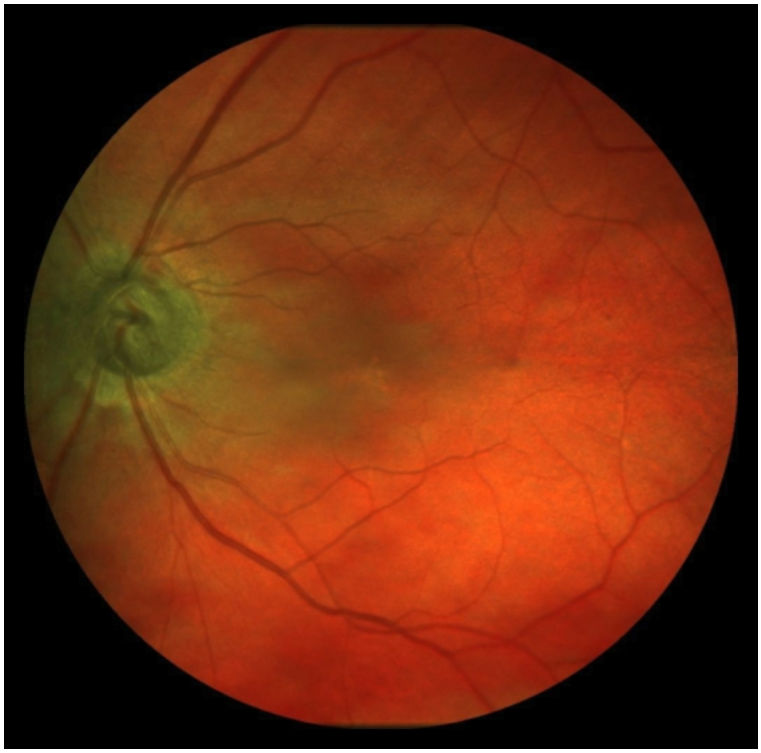


Figure 3.10: Example of one image from iostar dataset

Figure 3.11: Example of one image from drionsdb dataset

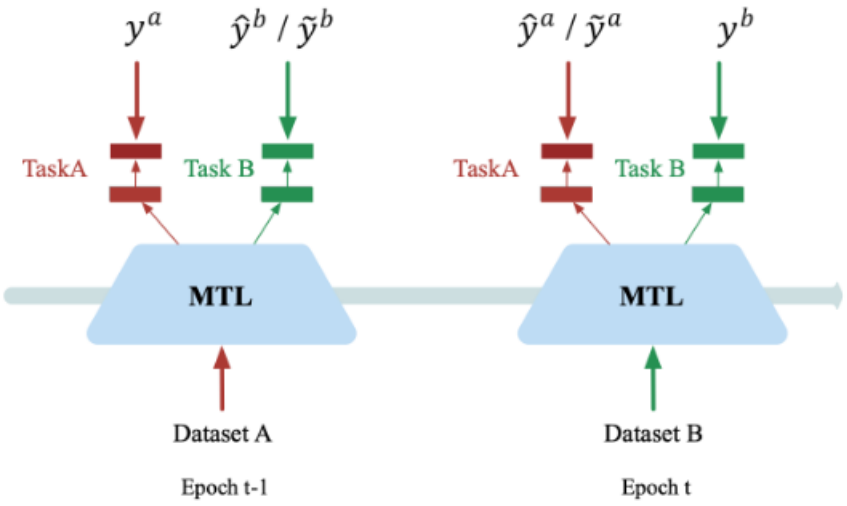


Figure 3.12: Alternating training strategy for Multi-task learning with disjoint datasets. Only two epochs are show in this figure, the left part is used on dataset A in epoch $t-1$, and the right part is used on dataset B in epoch t [35].

3.6.2 Multi-task baselines

Baseline Multi-task approach

Most of the MTL approaches concerning the segmentation problems are based on working with datasets containing the ground truth for both of the tasks, but in the medical field, it is not always the case to find datasets that combine all the tasks. This is why in our project we focused on working on an approach that works with disjoint datasets.

In this case, a naive approach would be to train each task for a complete epoch and then switch to the other task, but as we know, the model will have a hard time keeping the information from the previous task. In other words, in each epoch the network will forget the previous task, which will make it difficult for the model to converge.

That is why we explored from the literature an approach that facilitates the training on disjoint datasets and keeps track of all the tasks. The idea, as it is shown in figure 3.12 is to alternate the training of the tasks in each epoch. In other words, we train for the dataset A in epoch t with the ground truth of the dataset A, and for the ground truth of the second task, we take the predictor from the last epoch and use it to generate a ground truth in dataset A. In epoch t-1, we alternate and do the same but working on dataset B. The loss in this case is the sum of the two losses for each task.

In a more technical sense, each image I_i^a from dataset A in epoch t have a ground truth y^a but no ground truth for task B, so we use the predictor of epoch t-1 $p_{t-1}^b(.)$ and apply it to all the images from dataset A to generate ground truth for task B, which we call $\tilde{y}^b$, the weights update is done for each batch.

Experiments

To explore this approach further, we decided to focus on 1 model and 1 protocol, which are, respectively, Unet and the resizing 768 protocol, as we needed to have a protocol that works for both of the tasks at the same time. We prepared four experiments, which are :

1. Experiment 1 : in this experiment, we aim to train both of the tasks with a balanced approach. Each of the tasks will have a dataset with the same number of images. The quality of the images remains the same for each of the tasks.

2. Experiment 2 : in this experiment, we try to test if an unbalanced dataset would affect the training, so we take one dataset from each task with a different number of images and we see if it affects the results of the task with the small dataset, and once again, the quality of the images remains the same for each of the tasks.

3. Experiment 3 : in this experiment, we tried to explore the effect of having different quality of image in one of the tasks. We kept the balancing of data by using 3 datasets for vessels and one dataset for optic-disc.

4. Experiment 4 : in this experiment we tried to explore the effect of having a different quality of image in one of the tasks, but with an imbalanced set of datasets.

In this part, we are using a U-net with a different configuration to facilitate the implementation. The performance of this U-net is a bit lower than the previous one, but the goal here was to explore the MTL approach and analyze how we can present a generalizable approach.

In those experiments we explored 4 paths. The first one is to have two datasets with the same number of images for each task, so here we used Iostar-disc and drive respectively for optic-disc and vessels (20 images for vessels / 20 images for optic-disc). .The second one was to see if the imbalance of the data affected the performance of one of the tasks, so we used drive and drionsdb (20 images for vessels / 60 images for optic disc). In the third, we tried to keep the same number of images in both tasks but with different quality of images, so we used drive, hrf, and iostar for vessels and drionsdb for optic-disc (55 for vessels/ 60 for optic-disc). Finally, we tried an imbalance protocol with different image quality, so we used drive, hrf for vessels , and drionsdb for optic-disc(35 for vessels / 60 for optic-disc). More experiments should be done to test on all the datasets in order to explore the results more, but in our case, we tried to keep a small number of tests in order to analyze each experiment in all aspects. The results of these experiments are presented in tables 3.11 and 3.12. Tables 3.13 3.14 present the cross database results to evaluate the generalization for both vessels and optic-disc using the experiments discussed earlier. In this part, we are using a U-net with a different configuration to facilitate the implementation. The performance of this U-net is a bit lower than the previous one, but the goal here was to explore the MTL approach and analyze how we can present a generalizable approach.

Experience / Dataset	Drive	hrf	iostar
Experience 1	0.774	X	X
Experience 2	0.763	X	X
Experience 3	0.776	0.712	0.782
Experience 4	0.778	0.689	X
Single-task Unet with new configuration	0.802	0.794	0.803

Table 3.11: Baseline results for vessels with mtl approach

Experience / Dataset	Iostar-disc	Drionsdb
Experience 1	0.878	X
Experience 2	X	0.956
Experience 3	X	0.825
Experience 4	X	0.939
Single-task Unet with new configuration	0.923	0.963

Table 3.12: Baseline results for optic-disc with mtl approach

Experience / Dataset	Drive	hrf	iostar	chasedb	stare
Experience 1	**0.774**	0.636	0.685	0.627	0.717
Experience 2	**0.763**	0.717	0.654	0.627	0.717
Experience 3	**0.776**	**0.712**	**0.782**	0.724	0.731
Experience 4	**0.778**	**0.689**	0.747	0.683	0.740
Single-task Unet with new configuration DRIVE	**0.802**	0.721	0.786	0.754	0.793
Single-task Unet with new configuration HRF	0.693	**0.794**	0.679	0.593	0.725
Single-task Unet with new configuration IOSTAR	0.748	0.718	**0.803**	0.767	0.763

Table 3.13: cross database results for vessels with mtl approach

Experience / Dataset	Drionsdb	Iostar-disc	Rimoner3	Refuge-disc	Drishtigs1
Experience 1	0.370	**0.878**	0.302	0.093	0.322
Experience 2	**0.956**	0.559	0.453	0.675	0.857
Experience 3	**0.825**	0.272	0.527	0.344	0.747
Experience 4	**0.939**	0.370	0.509	0.856	0.892
Single-task Unet with new configuration Drionsdb	**0.963**	0.454	0.858	0.920	0.961
Single-task Unet with new configuration Iostar-disc	0.146	**0.923**	0.267	0.262	0.353

Table 3.14: cross database results for optic disc with mtl approach

Chapter 4

Analysis

This chapter presents an analysis of the baseline results and a proposition of two approaches to improve the results. The first proposition focuses on an internal change of the learning that consists of training the model using batch learning with an update of the weights at the end of each epoch. The second proposition consists of a small change to the proposed baseline architecture. Instead of taking the predictor of the previous epoch for the non-annotated task, we take the best predictor from all the previous epochs.

4.1 Analysis of the Multi-task baseline results

After collecting the baseline results, it was mandatory to analyze why the results didn't achieve a better generalization compared to single-task baselines. To do this, we generated for each experience a set of plots to see how the losses evolve over the 1000 epochs. In this analysis, we focus on five different plots. The first one represents the combined average training loss for all the tasks. The second one is the average validation loss for each task. The third plot corresponds to a separated average training loss for each task, in this plot, as we need to compare the evolution of the loss of each task, it was mandatory to have a plot easy to read. For this, we only take the loss where the model trained on the dataset with the real ground truth, and we duplicate this loss in the next epoch where the model trained on the other task. The last two plots corresponds to a comparaison between the Multi-task and single-task training loss on the same task.

In figure 4.1.a, we can see that training loss is noisy, until the last epochs where it becomes a bit smoother. The loss here presents the combined average for both tasks, as we train each epoch for a specific dataset with the real ground truth, and we train the other task with the predictor of the last epoch. The noise appear until the predictors start giving good annotation for the non annotated dataset as illustrated in figure 4.1.c. The validation part was always done separately using the datasets with the corresponding ground truths so that we can get the best model that provides the most accurate results on the well annotated datasets, we can notice that until the epoch 800 the validation losses were noisy as we were switching the tasks in each epoch as depicted in figure 4.1.b. Finally, we compared the average training loss for each task with the one we got in the single task approach, and we can remark that the model only converges to a certain point for both of the tasks, which shows that each the of the tasks affects each other and block the convergence for both of them as we can see in figure 4.1-d and figure 4.1-e.

In the second experience, we have two imbalanced datasets (60 images for optic-disc, 20 images for vessels). The quality of the image in drionsdb is different from iostar-disc, however we try to analyze the evolution of the loss to see if having more data in one task will have a negative effect on the performance. In figure 4.2.a, we can see that the loss is still noisy even in the end of the training, which means that the predictor on the non annotated task doesn't give a stable ground truth for the corresponding task. In figure 4.2.b, we can see that the validation loss for the optic-disc is below the one for the vessel task, and we can see that for the vessel task the loss didn't change a lot over all the epochs, which is also clear in figure 4.2.c where the vessel loss is not converging. This can be due to the imbalance of the datasets as in the model we update the weights in each batch of 4, so there are more updates when training the optic-disc compared to when training the vessel task. It is also clear in figure 4.2.d that the training loss for the vessel task is far away from the one we have in the single-task model, however the training loss for optic-disc is converging and is almost at the same level of that one in the single-task model as presented in the figure 4.2.e.

In the third experience, we explore the case of having a balanced dataset but with a different quality of the images. In figure 4.3.a we have a plot with more noise in the optic-disc task compared to the previous experience, this could be due to the optic-disc predictor has to predict the disc in the eye fundus image for all the type of the images, which means that the model can sometimes predict a good disc for the images of drive and give bad performance on one of the other vessels dataset. The noise is also clear in figure 4.3.b and

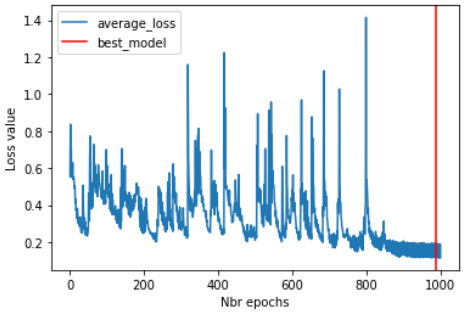
figure 4.3.c, which means that having different quality of images using our approach tends to affect the model, especially when we deal with updating the weight of the model after each batch. In figure 4.3.d and figure 4.3.e we can see once again that the MTL loss for each task fail to achieve the performance of the single task model.

In the fourth experience, we explore the case of having an imbalanced dataset with different quality of the images. We can observe in all the plots of 4.4 similar patterns as the previous experience, which is logical as the only difference between the two experience is the imbalance of the number of the images between the two tasks. On the other hand, baselines results in table 3.14 , and table 3.13 presents better results in optic-disc and almost similar performance in vessels. This could be due to the model weights that are updating more times when training the optic-disc task using the real annotation of the images.

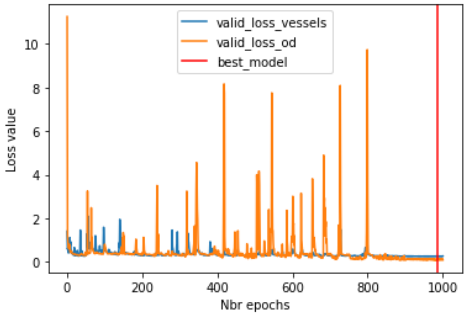
Finally, after analyzing the plots of the baseline method, we can see that there was two important problems.

1. The first problem consists of the convergence of the loss, we can see that in all the experiences both of the task failed to converge to a similar loss as the one in the single task model.

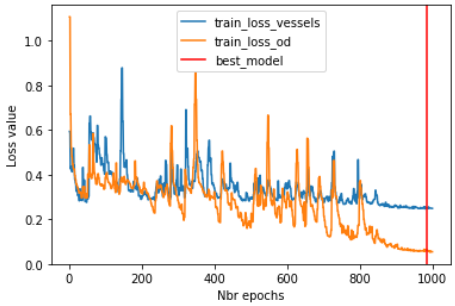
2. The second problem consists of the noise generated in the losses, we can clearly see that the loss is having a hard time to converge as it keeps going up and down between the epoches.



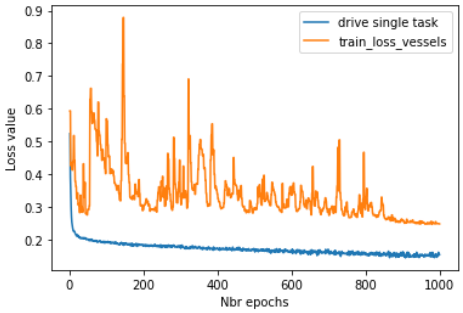
(a) Average training loss for both of the tasks



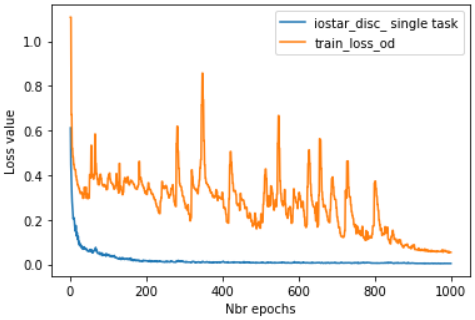
(b) Separated average validation loss



(c) Separated average training loss

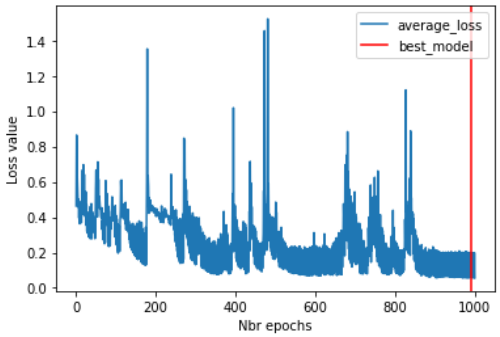


(d) Comparison between single-task and Multi-task average training loss for drive dataset

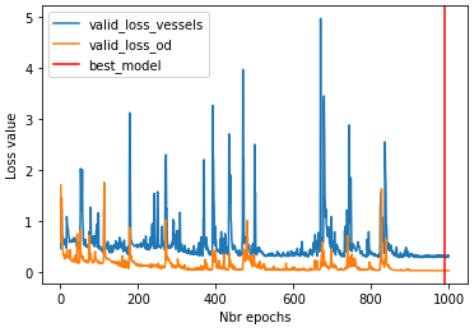


(e) Comparison between single-task and Multi-task average training loss for iostar-disc dataset

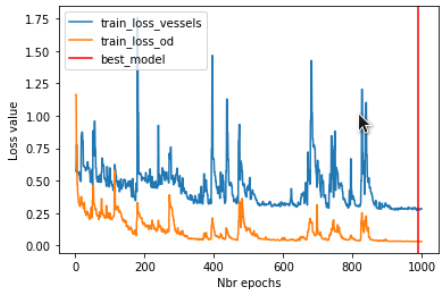
Figure 4.1: Experience 1 plots



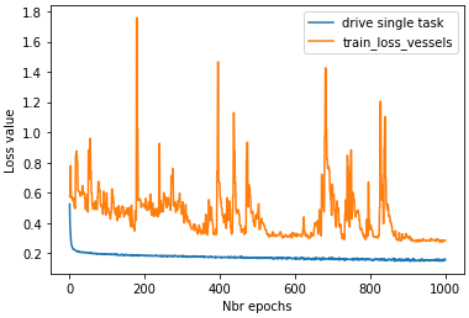
(a) Average training loss for both of the tasks



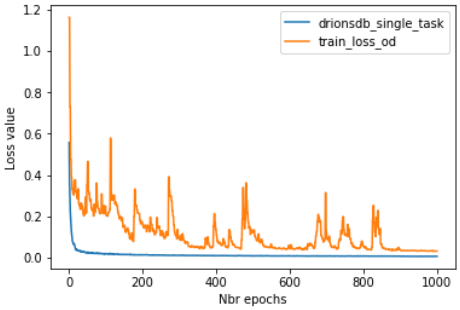
(b) Separated average validation loss



(c) Separated average training loss

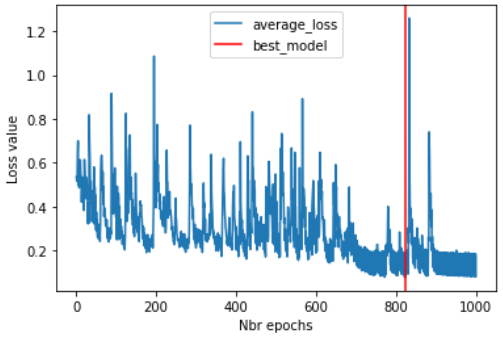


(d) Comparison between single-task and Multi-task average training loss for drive dataset

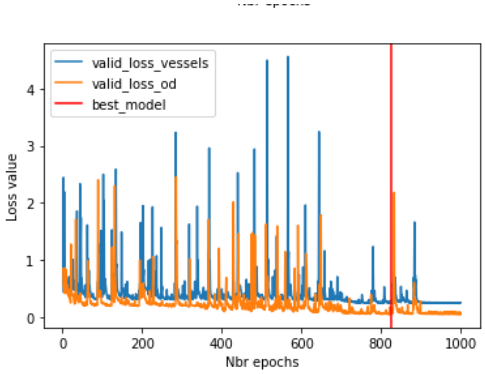


(e) Comparison between single-task and Multi-task average training loss for drionsdb dataset

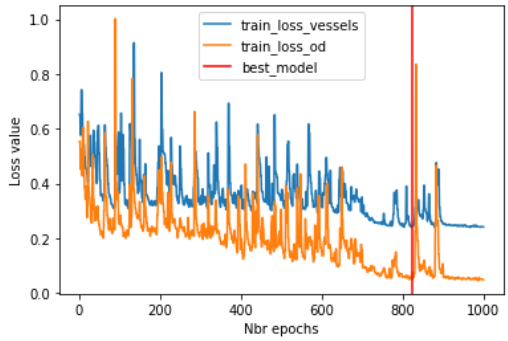
Figure 4.2: Experience 2 plots



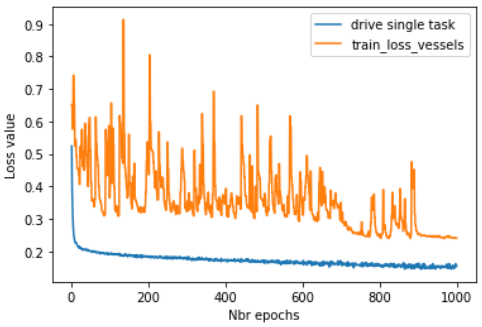
(a) Average training loss for both of the tasks



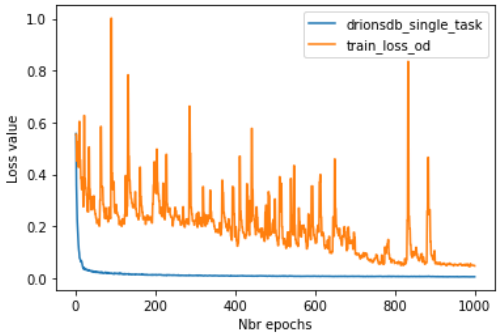
(b) Separated average validation loss



(c) Separated average training loss

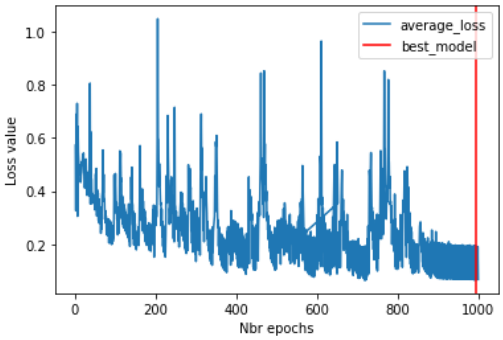


(d) Comparison between single-task and Multi-task average training loss for drive dataset

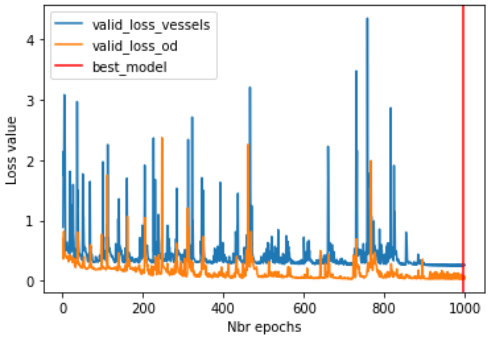


(e) Comparison between single-task and Multi-task average training loss for drionsdb dataset

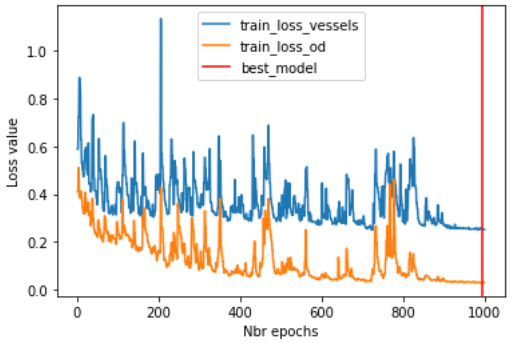
Figure 4.3: Experience 3 plots



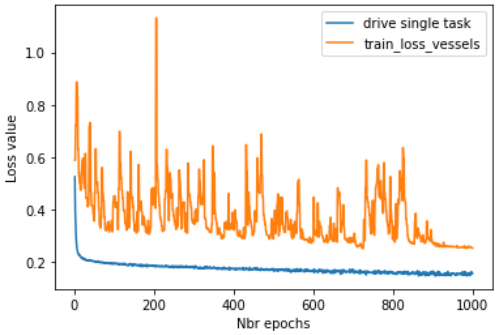
(a) Average training loss for both of the tasks



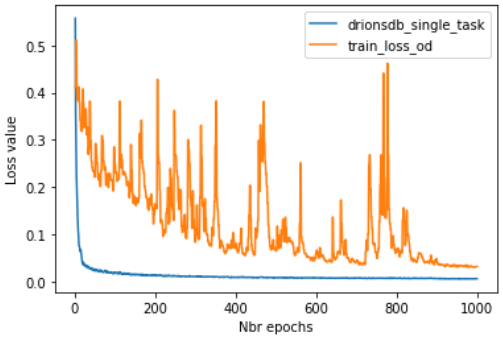
(b) Separated average validation loss



(c) Separated average training loss



(d) Comparison between single-task and Multi-task average training loss for drive dataset



(e) Comparison between single-task and Multi-task average training loss for drionsdb dataset

Figure 4.4: Experience 4 plots

Experience / Dataset	Drive	hrf	iostar
Experience 1	0.801	X	X
Experience 2	0.800	X	X
Experience 3	0.796	0.745	0.812
Experience 4	0.797	0.740	X
Single-task Unet with new configuration	0.802	0.794	0.803

Table 4.1: Results for vessels with gradient accumulation MTL approach

Experience / Dataset	Iostar-disc	Drionsdb
Experience 1	0.911	X
Experience 2	X	0.935
Experience 3	X	0.931
Experience 4	X	0.917
Single-task Unet with new configuration	0.923	0.963

Table 4.2: Results for optic-disc with gradient accumulation MTL approach

4.2 A new learning approach with gradient accumulation

In the previous section, we analyzed the baseline results and we saw that the validation losses in most of the cases was noisy, especially when we dealed with imbalance datasets. This was due to an update of the weight at each batch which makes the model forget the non annotated task. In order to avoid this much noise, we tried to implement the same approach with gradient accumulation, which means that we take the average loss of all the batches and we update the weight at the end of the epoch. This can also be useful when dealing with imbalanced datasets as we will have the same number of updates for each task.

In the first experiment, we can notice from figure 4.5.a and figure 4.5.c that the training loss is converging into a certain point for each of the tasks. The noise that appears in the combined training loss plot is due to the averaged part with the predictor of the non annotated

Experience / Dataset	Drive	hrf	iostar	chasedb	stare
Experience 1	**0.801**	0.717	0.654	0.694	0.753
Experience 2	**0.800**	0.696	0.766	0.665	0.766
Experience 3	**0.796**	**0.745**	**0.812**	0.754	0.715
Experience 4	**0.797**	**0.740**	0.782	0.723	0.722
Single-task Unet with new configuration DRIVE	**0.802**	0.721	0.786	0.754	0.793
Single-task Unet with new configuration HRF	0.693	**0.794**	0.679	0.593	0.725
Single-task Unet with new configuration IOSTAR	0.748	0.718	**0.803**	0.767	0.763

Table 4.3: Cross database results for vessels with gradient accumulation MTL approach

Experience / Dataset	Drionsdb	Iostar-disc	Rimoner3	Refuge-disc	Drishtigs1
Experience 1	0.071	**0.911**	0.081	0.093	0.067
Experience 2	**0.935**	0.625	0.709	0.871	0.921
Experience 3	**0.931**	0.057	0.695	0.900	0.769
Experience 4	**0.917**	0.429	0.405	0.915	0.775
Single-task Unet with new configuration Drionsdb	**0.963**	0.454	0.858	0.920	0.961
Single-task Unet with new configuration Iostar-disc	0.146	**0.923**	0.267	0.262	0.353

Table 4.4: Cross database results for optic disc with gradient accumulation MTL approach

task. In figure 4.5.b we can see that the validation is following the same curve as the training loss and that the curve is smoother than the one we had in the baseline approach. We can also see that the training loss is still going up and down between the epoches in figure 4.5.a. However, we can see that the training loss for each task separately is following a smoother curve as shown in figure 4.5.c. In this experience we can also see that the training curve for optic disc is almost reaching the loss with the single task model, but the one for the vessels is still far from it as depicted in 4.5.d and figure 4.5.e. The performance concerning this experience has given a better F1 score compared to the one from the baselines concerning the vessels segmentation as it's clearly illustrated in table 4.3, but it was not the case in the optic disc task in the term of generalization see table 4.4. Although the model gives better performance when using the test set of the same dataset, the generalization was really weak. As explained in the previous chapter, this could be due to the images used in iostar database, where the optic-disc is a bit different from the discs in the other datasets.

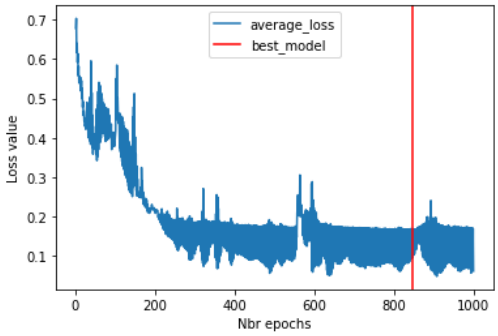
In the second experiment, we observe from figure 4.6 the same pattern in the previous experiment where the plots are smoother compared to the baseline method. This time, the results shows better generalization in both tasks as we can see in tables 4.3, 4.4. We can also see that the results in this experiment are a bit comparable with the one from the single tasks, although we are still facing the problem of the loss convergence especially in the vessel task as it's shown in figure 4.6.d, 4.6.e. It is clear from figure 4.6.b, 4.6.c that of both the training and validation loss reach a level of convergence and stops decreasing for each task.

In the third experiment, we can see again from tables 4.3, 4.4 that the performance for the vessels task has more or less improved as we gave more images to train, however we can clearly see that in term of generalization in the optic-disc task, the performance has decreased compared to the second experiment. Furthermore, we notice a big improvement of the performance in optic-disc task compared to the baseline method and a small improvement for the vessels task. As explained in the analysis of the baseline results, the decrease of performance compared to the previous experiment is due to the model that has a hard time to predict the disc for all the non annotated datasets. Using the gradient accumulation approach, we can remark from figure 4.7 that the plots are smoother once again compared to the baseline plots, which explains the improvement of the performance.

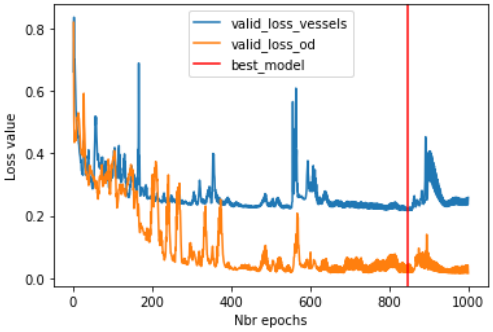
In the fourth experiment, we can observe from tables 4.3 4.4, that the results for both of the tasks are more or less the same with the previous experiments, except in the case of iostar-disc and rimoner3 where both of the datasets are a bit different from the other datasets. Comparing these results with baseline method, we can clearly see that the performance has increased for the case of vessels and decreased for the optic-disc task, which is logical as this experiment present an imblacement of the datasets. Hence, with this method we only update the weights at the end of the epoch which means we have less updates for the optic

disc-task compared to the baseline method. Analyzing the plots in figure 4.8 we can clearly see that the plots follows the same patterns as experiment 3, and also present smoother plots compared with the baseline plots.

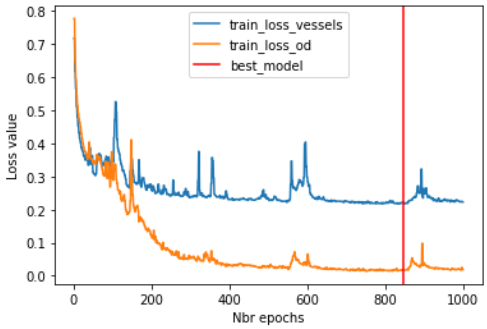
Looking at the plots of all the experiments, we can see that with this approach we were able to remove the noise from the loss curves, which shows that the loss converge more easily compared to the baseline method. However, looking at the performance of the experiments, we observed that in experiment 2 and experiment 4 which present an imbalance of the datasets, there was a clear improvement in experiment 2 for the optic-disc task compared with the baseline method, but it was not the case for experiment 4. With the gradient accumulation method the update of the weights is done once at the end of each epoch, so in experiment 2 we can see that the baseline model gave a better performance only in the dataset that was used in training, as the model presented a lot of noise the generalization performance was worse than the performance with the gradient accumulation method. In experiment 4, the loss were smoother so when decreasing the number of updates it affected the performance of the optic-disc task. However, in both experiments we saw an increasement of the performance for the vessel task, so we can conclude that the model with the gradient accumulation is more adapted for imbalanced datasets.



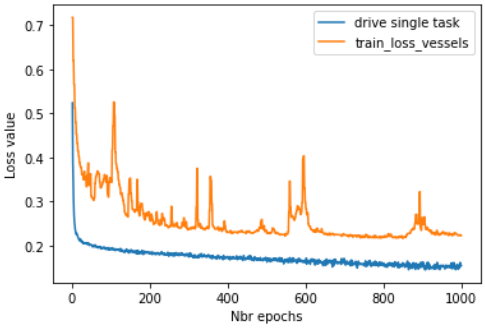
(a) Average training loss for both of the tasks



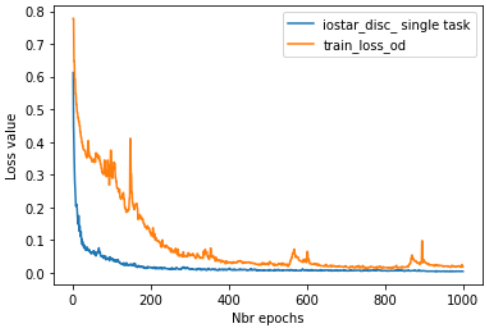
(b) Separated average validation loss



(c) Separated average training loss

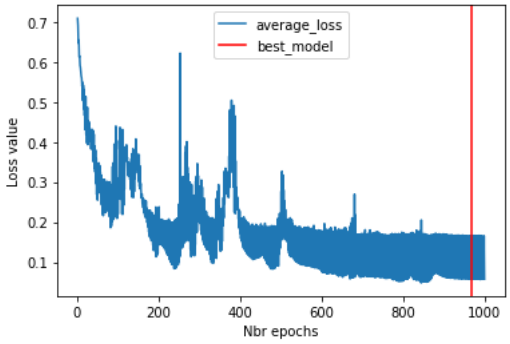


(d) Comparison between single-task and Multi-task average training loss for drive dataset

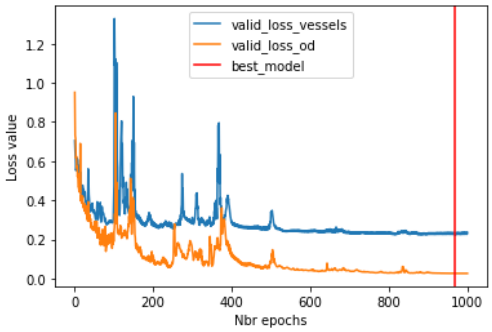


(e) Comparison between single-task and Multi-task average training loss for iostar-disc dataset

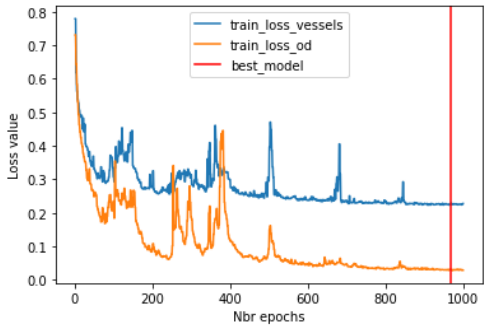
Figure 4.5: Experience 1 plots with gradient accumulation



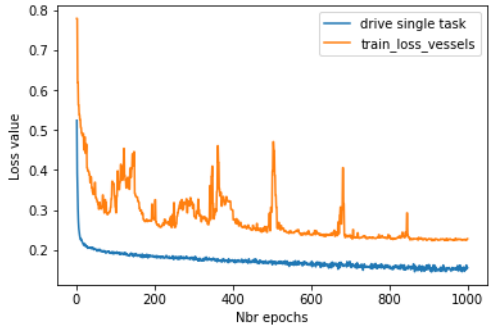
(a) Average training loss for both of the tasks



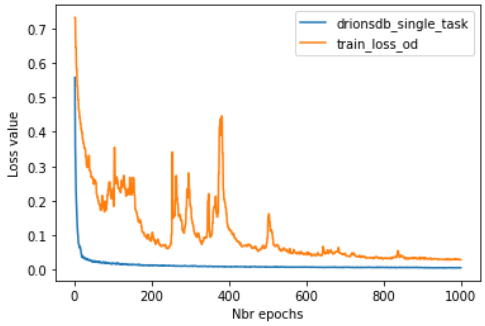
(b) Separated average validation loss



(c) Separated average training loss

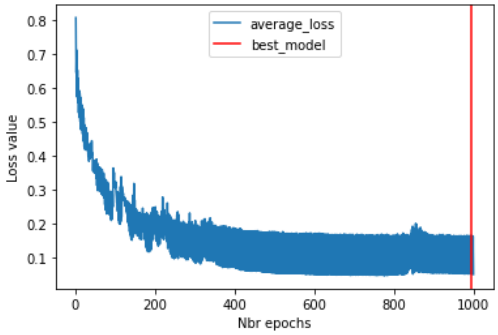


(d) Comparison between single-task and Multi-task average training loss for drive dataset

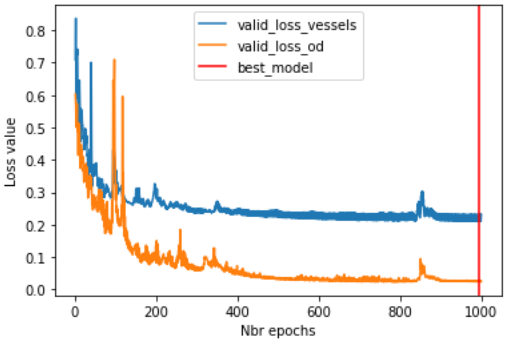


(e) Comparison between single-task and Multi-task average training loss for drionsdb dataset

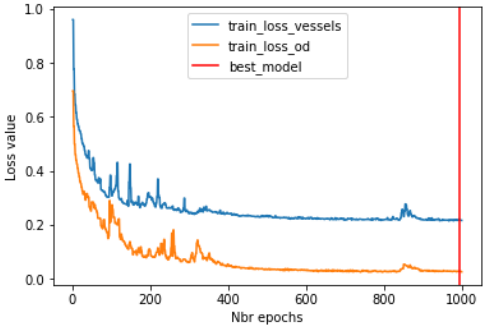
Figure 4.6: Experience 2 plots with gradient accumulation



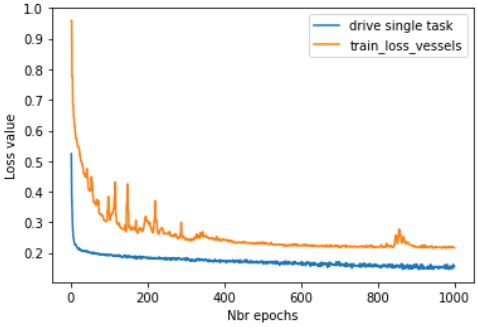
(a) Average training loss for both of the tasks



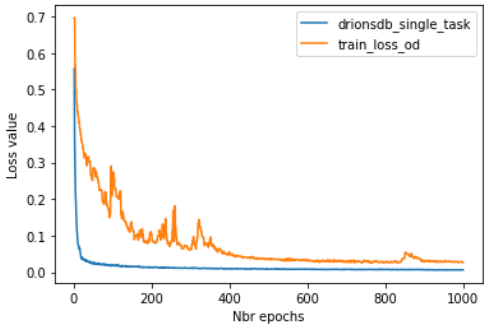
(b) Separated average validation loss



(c) Separated average training loss

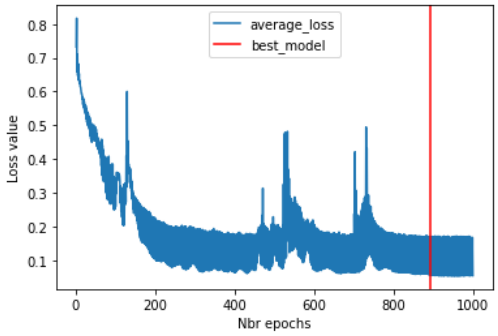


(d) Comparison between single-task and Multi-task average training loss for drive dataset

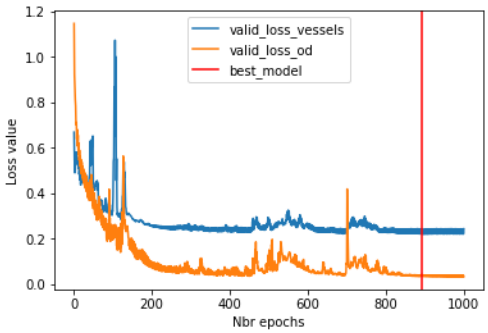


(e) Comparison between single-task and Multi-task average training loss for drionsdb dataset

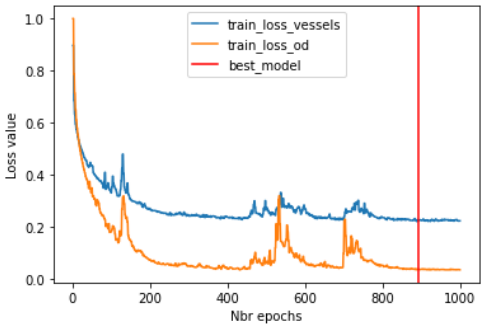
Figure 4.7: Experience 3 plots with gradient accumulation



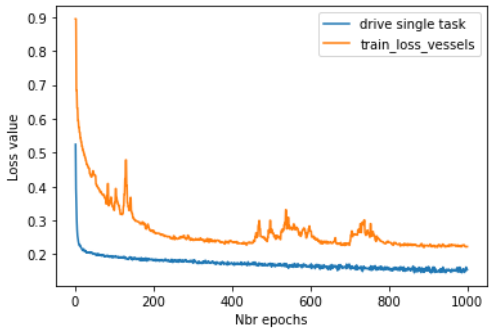
(a) Average training loss for both of the tasks



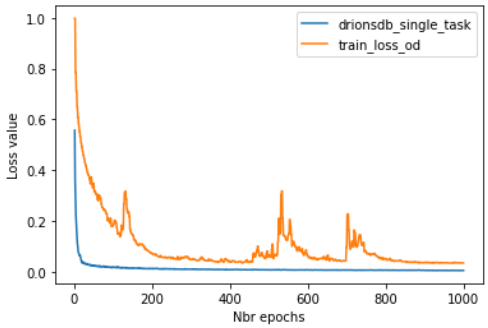
(b) Separated average validation loss



(c) Separated average training loss



(d) Comparison between single-task and Multi-task average training loss for drive dataset



(e) Comparison between single-task and Multi-task average training loss for drionsdb dataset

Figure 4.8: Experience 4 plots with gradient accumulation

4.3 Another approach for Multi-task learning with disjoint datasets

We saw in the previous parts that the models failed to converge to the level of the single task approach, and we saw that this could be due to the effect of the bad ground truth generated from the predictor of the previous epoch. We proposed therefore a method that try to avoid this problem. Instead of taking the predictor of the previous epoch, we saved the predictor of the epoch which gave the best validation loss for each task, which means that we take the best predictor for each task and use it as a generator of the ground truth. To compare the results, we used the same set of experiment with both the approaches that was described before (Batch update, gradient accumulation). As we can see from tables 4.5, 4.6, the results for both optic-disc and and vessels are more or less similar to the previous experience, and we can also see from the figure 4.9 that the value of the training loss for each task in all the experiments converge to a higher loss compared with the previous method. In this case, instead of analyzing all the different plot, we focus on the separated training tasks, in order to see the effect of this method on the development of the the training loss. Concerning the experiments with the batch update approach, we can still remark the presence of the noise, but this time we notice some big changes in the losses in certain epoches, which makes it harder for the model to converge to a small loss. For the experiments with gradient accumulation, as we said before it helps to have a smoother curve in the loss, which is the case for our current experiments. However, we can clearly see the same problem of the sudden loss change which affect the converegence of the loss. In both approaches used with this new MTL method, we can remark that the best epoch model for each experiment is at the middle of training except for experiment 2 with batch update, which was not the case in the previous method, as the best model wa always at the last epochs. The model in this method, takes the best predictor for each task to generate the ground truth for the non annotated task, and sometimes the best predictor is taken from epochs that are far away from the current epoch, which means that the weights are a lot different compared to the current epoch, which creates the sudden change of the loss as the model is trying to optimize the combined loss. This behaviour of the loss was expected, however we hypothesized that using this method we would get a better combined loss, as we would generate better ground truths for the non annotated task. This was not the case in our experiments, which means that we generated more noise in the plots without solving the problem of the convergence of the loss. Finally, we can say that this approach is less efficient than the model which takes the predictor of the previous epoch.

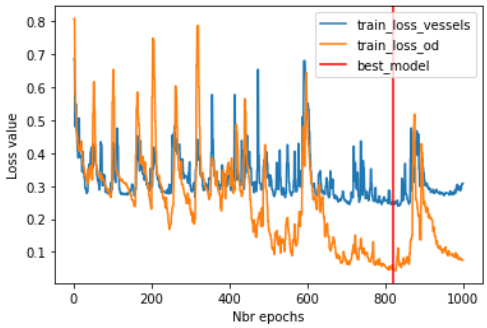
Analyzing the baseline plots, we generated two problems, the first one was the noise created by the batch update method. This problem was partially solved using the gradient accumulation approach that generated smoother curves in the loss. The second problem was the convergence of the loss into an equivalent loss of the single task model. This method was tried to face this problem, however, it only created more noise in the loss, and made it harder for the model to converge.

Experience / Dataset	Drive	hrf	iostar
Experience 1 (batch update)	0.785	X	X
Experience 1 (gradient accumulation)	0.785	X	X
Experience 2 (batch update)	0.751	X	X
Experience 2 (gradient accumulation)	0.783	X	X
Experience 3 (batch update)	0.766	0.745	0.812
Experience 3 (gradient accumulation)	0.789	0.745	0.812
Experience 4 (batch update)	0.783	0.740	X
Experience 4 (gradient accumulation)	0.790	0.740	X
Single-task Unet with new configuration	0.802	0.794	0.803

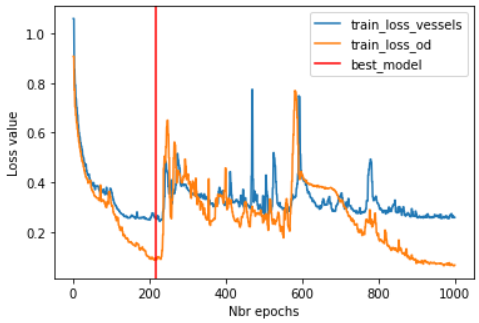
Table 4.5: Results for vessels with new mtl approach

Experience / Dataset	Iostar-disc	Drionsdb
Experience 1 (batch update)	0.916	X
Experience 1 (gradient accumulation)	0.905	X
Experience 2 (batch update)	X	0.932
Experience 2 (gradient accumulation)	X	0.920
Experience 3 (batch update)	X	0.900
Experience 3 (gradient accumulation)	X	0.913
Experience 4 (batch update)	X	0.930
Experience 4 (gradient accumulation)	X	0.929
Single-task Unet with new configuration	0.923	0.963

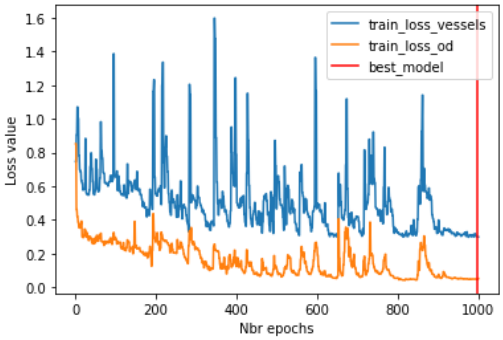
Table 4.6: Results for optic-disc with gradient accumulation mtl approach



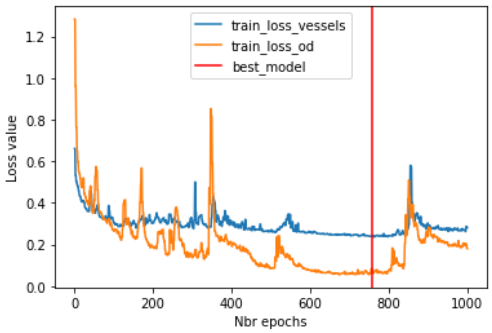
(a) Experiment 1 with batch update



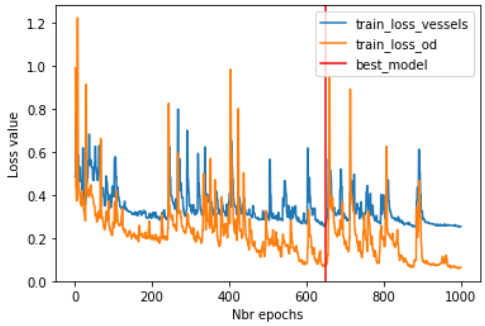
(b) Experiment 1 with gradient accumulation



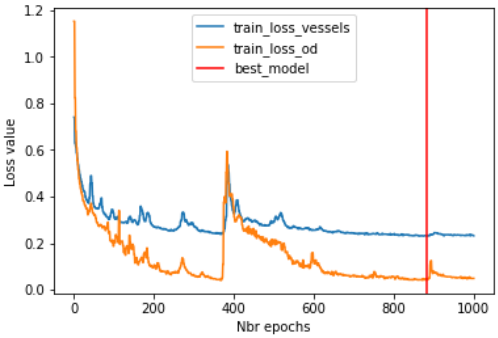
(c) Experiment 2 with batch update



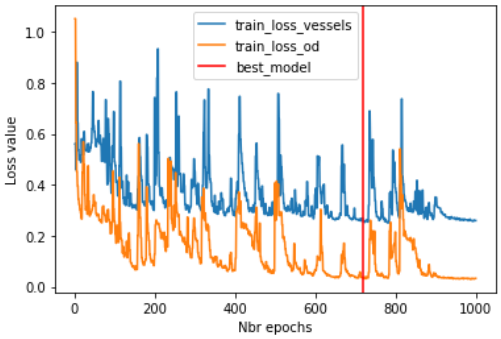
(d) Experiment 2 with gradient accumulation



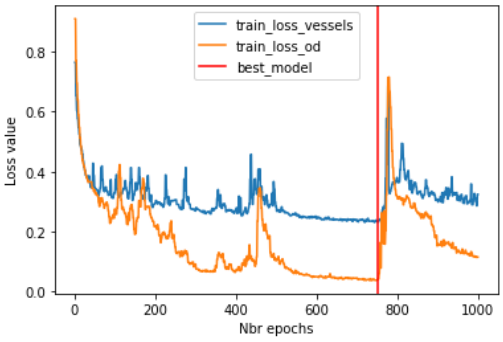
(e) Experiment 3 with batch update



(f) Experiment 3 with gradient accumulation



(g) Experiment 4 with batch update



(h) Experiment 4 with gradient accumulation

Figure 4.9: Separeted training tasks plots for all the experiments with the new mtl approach

Conclusion

In this work, it has been attempted to compare and analyze a Multi-task approach for disjoints datasets. We hypothesized that an MTL framework would generate a better generalization than a single task approach. However, the results have shown a decrease in performance. Our MTL model has failed to give a better result in the images with the same quality as the trained datasets and even worse results in datasets with different quality of images, which means that the model failed to generalize, and we could not verify our hypothesis. Four experiments were tested to understand the effect of the imbalance and the different quality of the data effects on training a Multi-task model in disjoint datasets. Analyzing the results, we saw that the baseline method in the four experiments presented two common problems at the training level. The first was the noise of the losses caused by the multiple weight updates, making it harder for the model to converge. The second problem was that the model failed to converge to the same level as the single-task approach. Our hypothesis for this problem was that the model didn't provide good ground truth for the non-annotated task. The next step was to find solutions for both of the problems, so we proposed an approach with gradient accumulation to solve the first problem, this method's goal was to decrease the noise of the losses by only updating the weight once at each epoch, and we remarked that the approach succeeded to decrease the noise and give better results than the one from the baseline method. The second proposed approach was to take the best epoch for each task to generate the ground truth for the non-annotated task. Still, after analyzing the results and the plots, we only remarked that the model became noisier and even stopped in higher loss compared to the baseline method. This analysis showed us that convergence failure was not due to the bad ground truth.

After looking again at all the plots, we remarked that each task badly affects the convergence of the other task, which made us think that the model failed to converge as the task are a bit non-related. A solution to this problem would be to use the same gradient accumulation approach but with a different model with more specific layers for each task. In our model, most of the U-Net layers were shared between both tasks, which means that the model had a hard time converging. We hypothesize that adding more specific layers for each task would solve the convergence problem, and we can even get a more generalizable model compared to the single-task model. Another solution is to switch the model into a soft parameter sharing. Even though the model would be less scalable, the loss will surely be able to converge better compared to our current hard-parameter sharing model with few specific layers.

Bibliography

[1] Samaneh Abbasi-Sureshjani et al. "Biologically-Inspired Supervised Vasculature Segmentation in SLO Retinal Fundus Images". In: *Image Analysis and Recognition*. Ed. by Mohamed Kamel and Aurélio Campilho. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 325–334. ISBN: 9783319208015. DOI: 10.1007/978-3-319-20801-5_35.

[2] M. Abràmoff et al. "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning." In: *Investigative ophthalmology & visual science* 57 13 (2016), pp. 5200–5206.

[3] *Age-related macular degeneration (AMD) | Swiss Medical Network*. URL: https://www.swissmedical.net/en/our-specialties/ophthalmic-surgery-age-related-macular-degeneration-amd/ (visited on 06/28/2021).

[4] *An Overview of Multi-Task Learning for Deep Learning*. Sebastian Ruder. May 29, 2017. URL: https://ruder.io/multi-task/ (visited on 05/16/2022).

[5] Maryam Badar, M. Shahzad, and M. Fraz. "Simultaneous Segmentation of Multiple Retinal Pathologies Using Fully Convolutional Deep Neural Network". In: *MIUA*. 2018.

[6] Baidaa Al-Bander et al. "Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc". In: *Biomed. Signal Process. Control.* 40 (2018), pp. 91–101.

[7] Braithwaite T Bourne RRA Flaxman SR. "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis". In: *The Lancet. Global health* (). URL: https://pubmed.ncbi.nlm.nih.gov/28779882/.

[8] C. Brandl et al. "Features of Age-Related Macular Degeneration in the General Adults and Their Dependency on Age, Sex, and Smoking: Results from the German KORA Study". In: *PLoS ONE* 11 (2016).

[9] A. Budai et al. "Robust Vessel Segmentation in Fundus Images". In: *International Journal of Biomedical Imaging* 2013 (Dec. 12, 2013), e154860. ISSN: 1687-4188. DOI: 10.1155/2013/154860. URL: https://www.hindawi.com/journals/ijbi/2013/154860/ (visited on 07/10/2021).

[10] P. Burlina et al. "Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks". In: *JAMA Ophthalmology* 135 (2017), pp. 1170–1176.

[11] P. Burlina et al. “Detection of age-related macular degeneration via deep learning”. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (2016), pp. 184–188.

[12] Enrique J. Carmona et al. “Identification of the optic nerve head with genetic algorithms”. In: *Artificial intelligence in medicine* 43 3 (2008), pp. 243–59.

[13] Rich Caruana. “Multitask Learning”. In: *Machine Learning* 28.1 (July 1, 1997), pp. 41–75. ISSN: 1573-0565. DOI: 10.1023/A:1007379606734. URL: https://doi.org/10.1023/A:1007379606734 (visited on 05/15/2022).

[14] A. Chakravarty and J. Sivaswamy. “A Deep Learning based Joint Segmentation and Classification Framework for Glaucoma Assesment in Retinal Color Fundus Images”. In: *ArXiv* abs/1808.01355 (2018).

[15] *China Focus: AI beats human doctors in neuroimaging recognition contest*. URL: http://www.xinhuanet.com/english/2018-06/30/c_137292451.htm.

[16] Piotr Chudzik et al. “DISCERN: Generative Framework for Vessel Segmentation using Convolutional Neural Network and Visual Codebook”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018), pp. 5934–5937.

[17] T. Clemons et al. “National Eye Institute Visual Function Questionnaire in the Age-Related Eye Disease Study (AREDS): AREDS Report No. 10.” In: *Archives of ophthalmology* 121 2 (2003), pp. 211–7.

[18] Adrián Colomer, Jorge Igual, and Valery Naranjo. “Detection of Early Signs of Diabetic Retinopathy Based on Textural and Morphological Information in Fundus Images”. In: *Sensors (Basel, Switzerland)* 20.4 (Feb. 13, 2020), p. 1005. ISSN: 1424-8220. DOI: 10.3390/s20041005. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7071097/ (visited on 06/28/2021).

[19] *Common Eye Disorders and Diseases | CDC*. June 4, 2020. URL: https://www.cdc.gov/visionhealth/basics/ced/index.html (visited on 06/28/2021).

[20] Michael Crawshaw. “Multi-Task Learning with Deep Neural Networks: A Survey”. In: *arXiv:2009.09796 [cs, stat]* (Sept. 10, 2020). arXiv: 2009.09796. URL: http://arxiv.org/abs/2009.09796 (visited on 05/15/2022).

[21] E. Decencière et al. “FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE”. In: *Image Analysis & Stereology* 33 (2014), pp. 231–234.

[22] E. Decencière et al. “TeleOphta: Machine learning and image processing methods for teleophthalmology”. In: *Irbm* 34 (2013), pp. 196–203.

[23] *Detect Glaucoma Early To Protect Vision*. NIH News in Health. May 10, 2017. URL: https://newsinhealth.nih.gov/2015/01/detect-glaucoma-early-protect-vision (visited on 07/11/2021).

[24] *Diabetic retinopathy*. nhs.uk. Oct. 23, 2017. URL: https://www.nhs.uk/conditions/diabetic-retinopathy/ (visited on 06/28/2021).

[25] Venkata Gopal Edupuganti, Akshay Chawla, and Amit Kale. "Automatic Optic Disk and Cup Segmentation of Fundus Images Using Deep Learning". In: *2018 25th IEEE International Conference on Image Processing (ICIP)* (2018), pp. 2227–2231.

[26] Rolando Estrada et al. "Retinal Artery-Vein Classification via Topology Estimation". In: *IEEE Transactions on Medical Imaging* 34.12 (Dec. 2015), pp. 2518–2534. ISSN: 1558-254X. DOI: 10.1109/TMI.2015.2443117.

[27] Zhongwei Feng et al. "Deep Retinal Image Segmentation: A FCN-Based Architecture with Short and Long Skip Connections for Retinal Image Segmentation". In: *ICONIP*. 2017.

[28] Adrian Galdran et al. "The Little W-Net That Could: State-of-the-Art Retinal Vessel Segmentation with Minimalistic Models". In: *arXiv:2009.01907 [cs, eess]* (Sept. 3, 2020). arXiv: 2009.01907. URL: http://arxiv.org/abs/2009.01907 (visited on 05/16/2022).

[29] Gabriel García et al. "Detection of Diabetic Retinopathy Based on a Convolutional Neural Network Using Retinal Fundus Images". In: *ICANN*. 2017.

[30] Rishab Gargeya and Theodore Leng. "Automated Identification of Diabetic Retinopathy Using Deep Learning." In: *Ophthalmology* 124 7 (2017), pp. 962–969.

[31] Arun Govindaiah et al. "Deep convolutional neural network based screening and assessment of age-related macular degeneration from fundus images". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018), pp. 1525–1528.

[32] Felix Grassmann et al. "A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography." In: *Ophthalmology* 125 9 (2018), pp. 1410–1420.

[33] M. V. Grinsven et al. "Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images". In: *IEEE Transactions on Medical Imaging* 35 (2016), pp. 1273–1284.

[34] Sven Holm et al. "DR HAGIS-a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients". In: *Journal of Medical Imaging (Bellingham, Wash.)* 4.1 (Jan. 2017), p. 014503. ISSN: 2329-4302. DOI: 10.1117/1.JMI.4.1.014503.

[35] Yan Hong et al. "Beyond without Forgetting: Multi-Task Learning for Classification with Disjoint Datasets". In: *arXiv:2003.06746 [cs, stat]* (Mar. 14, 2020). arXiv: 2003.06746. URL: http://arxiv.org/abs/2003.06746 (visited on 05/16/2022).

[36] A. Hoover, V. Kouznetsova, and M. Goldbaum. "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response". In: *IEEE Transactions on Medical Imaging* 19 (2000), pp. 203–210.

[37] Arnaldo Horta et al. "A Hybrid Approach for Incorporating Deep Visual Features and Side Channel Information with Applications to AMD Detection". In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2017), pp. 716–720.

[38] http://fyra.io. *Early Detection of AMD Is Crucial to Preserve Vision*. Retina Today. URL: https://retinatoday.com/articles/2007-nov/1107_03-php (visited on 07/11/2021).

[39] Vladimir Iglovikov, Sergey Mushinskiy, and Vladimir Osin. "Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition". In: *arXiv:1706.06169 [cs]* (June 19, 2017). arXiv: 1706.06169. URL: http://arxiv.org/abs/1706.06169 (visited on 05/16/2022).

[40] Vladimir I. Iglovikov et al. "TernausNetV2: Fully Convolutional Network for Instance Segmentation". In: *arXiv:1806.00844 [cs]* (June 19, 2018). arXiv: 1806.00844. URL: http://arxiv.org/abs/1806.00844 (visited on 05/16/2022).

[41] M. Jamshidi et al. "Automatic Detection of the Optic Disc of the Retina: A Fast Method". In: *Journal of Medical Signals and Sensors* 6.1 (2016), pp. 57–63. ISSN: 2228-7477. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4786964/ (visited on 07/11/2021).

[42] Kaggle. "https://www.kaggle.com/c/diabetic-retinopathy-detection/data". In: 2019.

[43] T. Kauppi et al. "DIARETDB 0 : Evaluation Database and Methodology for Diabetic Retinopathy Algorithms". In: 2007.

[44] T. Kauppi et al. "The DIARETDB1 Diabetic Retinopathy Database and Evaluation Protocol". In: *BMVC*. 2007.

[45] P. Khojasteh et al. "Introducing a Novel Layer in Convolutional Neural Network for Automatic Identification of Diabetic Retinopathy". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018), pp. 5938–5941.

[46] Tim Laibacher, Tillman Weyde, and Sepehr Jalali. "M2U-Net: Effective and Efficient Retinal Vessel Segmentation for Resource-Constrained Environments". In: *arXiv:1811.07738 [cs]* (Apr. 23, 2019). arXiv: 1811.07738. URL: http://arxiv.org/abs/1811.07738 (visited on 05/16/2022).

[47] Carson K. Lam et al. "Retinal Lesion Detection With Deep Learning Using Image Patches". In: *Investigative Ophthalmology & Visual Science* 59 (2018), pp. 590–596.

[48] Gabriel Lepetit-Aimon, R. Duval, and F. Cheriet. "Large Receptive Field Fully Convolutional Network for Semantic Segmentation of Retinal Vasculature in Fundus Images". In: *COMPAY/OMIA@MICCAI*. 2018.

[49] Gilbert Lim et al. "Integrated Optic Disc and Cup Segmentation with Deep Learning". In: *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)* (2015), pp. 162–169.

[50] G. Lin et al. "Transforming Retinal Photographs to Entropy Images in Deep Learning to Improve Automated Detection for Diabetic Retinopathy". In: *Journal of Ophthalmology* 2018 (2018).

[51] Paweł Liskowski and K. Krawiec. "Segmenting Retinal Blood Vessels With Deep Neural Networks". In: *IEEE Transactions on Medical Imaging* 35 (2016), pp. 2369–2380.

[52] Wenao Ma et al. "Multi-task Neural Networks with Spatial Activation for Retinal Vessel Segmentation and Artery/Vein Classification". In: *ArXiv* abs/2007.09337 (2019).

[53] Debapriya Maji et al. "Deep neural network and random forest hybrid architecture for learning to detect retinal vessels in fundus images". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015), pp. 3029–3032.

[54] K. Maninis et al. "Deep Retinal Image Understanding". In: *ArXiv* abs/1609.01103 (2016).

[55] Kevis-Kokitsi Maninis et al. "Deep Retinal Image Understanding". In: *arXiv:1609.01103 [cs]* 9901 (2016), pp. 140–148. DOI: 10.1007/978-3-319-46723-8_17. arXiv: 1609.01103. URL: http://arxiv.org/abs/1609.01103 (visited on 05/16/2022).

[56] Langis Michaud and Pierre Forcier. "Prevalence of asymptomatic ocular conditions in subjects with refractive-based symptoms". In: *Journal of optometry* (2014). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4087174/.

[57] Anirban Mitra et al. "The region of interest localization for glaucoma analysis from retinal fundus image using deep learning". In: *Computer methods and programs in biomedicine* 165 (2018), pp. 25–35.

[58] Benjamin Mueller. *The Artificial Intelligence Act: A Quick Explainer*. Center for Data Innovation. May 4, 2021. URL: https://datainnovation.org/2021/05/the-artificial-intelligence-act-a-quick-explainer/ (visited on 06/29/2021).

[59] "Ophthalmic diagnosis using deep learning with fundus images – A critical review". In: *Artificial Intelligence in Medicine* 102 (Jan. 1, 2020), p. 101758. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2019.101758. URL: https://www.sciencedirect.com/science/article/pii/S0933365719305858 (visited on 06/28/2021).

[60] J. Orlando et al. "An ensemble deep learning based approach for red lesion detection in fundus images". In: *Computer methods and programs in biomedicine* 153 (2018), pp. 115–127.

[61] C. Owen et al. "Measuring retinal vessel tortuosity in 10-year-old children: validation of the Computer-Assisted Image Analysis of the Retina (CAIAR) program." In: *Investigative ophthalmology & visual science* 50 5 (2009), pp. 2004–10.

[62] Abhishek Pal, Manav Rajiv Moorthy, and A. Shahina. "G-Eyenet: A Convolutional Autoencoding Classifier Framework for the Detection of Glaucoma from Retinal Fundus Images". In: *2018 25th IEEE International Conference on Image Processing (ICIP)* (2018), pp. 2775–2779.

[63] Oscar J. Perdomo et al. "Glaucoma Diagnosis from Eye Fundus Images Based on Deep Morphometric Feature Estimation". In: *COMPAY/OMIA@MICCAI*. 2018.

[64] Luis Perez and Jason Wang. "The Effectiveness of Data Augmentation in Image Classification using Deep Learning". In: *arXiv:1712.04621 [cs]* (Dec. 13, 2017). arXiv: 1712.04621. URL: http://arxiv.org/abs/1712.04621 (visited on 05/16/2022).

[65] Clément Playout, R. Duval, and F. Cheriet. "A Novel Weakly Supervised Multitask Architecture for Retinal Lesions Segmentation on Fundus Images". In: *IEEE Transactions on Medical Imaging* 38 (2019), pp. 2434–2444.

[66] G. Quellec et al. "Deep image mining for diabetic retinopathy screening". In: *Medical Image Analysis* 39 (2017), pp. 178–193.

[67] Refuge. "5th MICCAI workshop on ophthalmic medical image analysis (OMIA)." In: 2018.

[68] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv:1505.04597 [cs]* (May 18, 2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597 (visited on 05/16/2022).

[69] Oindrila Saha, R. Sathish, and D. Sheet. "Learning with Multitask Adversaries using Weakly Labelled Data for Semantic Segmentation in Retinal Images". In: *MIDL*. 2019.

[70] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In: *PLoS ONE* 10.3 (Mar. 4, 2015), e0118432. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118432. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/ (visited on 05/16/2022).

[71] Mark Sandler et al. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *arXiv:1801.04381 [cs]* (Mar. 21, 2019). arXiv: 1801.04381. URL: http://arxiv.org/abs/1801.04381 (visited on 05/19/2022).

[72] A. Sevastopolsky. "Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network". In: *Pattern Recognition and Image Analysis* 27 (2017), pp. 618–624.

[73] Juan Shan and L. Li. "A Deep Learning Method for Microaneurysm Detection in Fundus Images". In: *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (2016), pp. 357–358.

[74] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv:1409.1556 [cs]* (Apr. 10, 2015). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556 (visited on 07/05/2021).

[75] D. Siva Sundhara Raja and S. Vasuki. "Automatic Detection of Blood Vessels in Retinal Images for Diabetic Retinopathy Diagnosis". In: *Computational and Mathematical Methods in Medicine* 2015 (Feb. 24, 2015), e419279. ISSN: 1748-670X. DOI: 10.1155/2015/419279. URL: https://www.hindawi.com/journals/cmmm/2015/419279/ (visited on 07/11/2021).

[76] J. Sivaswamy et al. "Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation". In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (2014), pp. 53–56.

[77] J. Staal et al. "Ridge-based vessel segmentation in color images of the retina". In: *IEEE Transactions on Medical Imaging* 23 (2004), pp. 501–509.

[78] *The Benefits of AI in Healthcare - Pros and Cons*. Aidoc. Sept. 7, 2020. URL: https://www.aidoc.com/blog/pros-cons-artificial-intelligence-in-healthcare/ (visited on 06/28/2021).

[79] Zhi Tu et al. "SUNet: A Lesion Regularized Model for Simultaneous Diabetic Retinopathy and Diabetic Macular Edema Grading". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020), pp. 1378–1382.

[80] *Vision: Definition, Parts Of The Eye & Eye Health*. Cleveland Clinic. URL: https://my.clevelandclinic.org/health/articles/21204-vision (visited on 06/28/2021).

[81] Thomas Walter and Jean-Claude Klein. "Segmentation of Color Fundus Images of the Human Retina: Detection of the Optic Disc and the Vascular Tree Using Morphological Techniques". In: *Medical Data Analysis*. Ed. by Jose Crespo, Victor Maojo, and Fernando Martin. Berlin, Heidelberg: Springer, 2001, pp. 282–287. ISBN: 9783540454977. DOI: 10.1007/3-540-45497-7_43.

[82] *What Is Macular Degeneration?* American Academy of Ophthalmology. Jan. 26, 2021. URL: https://www.aao.org/eye-health/diseases/amd-macular-degeneration (visited on 06/28/2021).

[83] *WHO/Europe is focusing on eye screening for people with diabetes*. URL: https://www.euro.who.int/en/health-topics/noncommunicable-diseases/diabetes/news/news/2020/11/whoeurope-is-focusing-on-eye-screening-for-people-with-diabetes (visited on 06/28/2021).

[84] Saining Xie and Zhuowen Tu. "Holistically-Nested Edge Detection". In: *arXiv:1504.06375 [cs]* (Oct. 3, 2015). arXiv: 1504.06375. URL: http://arxiv.org/abs/1504.06375 (visited on 05/16/2022).

[85] Jiong Zhang et al. "Robust Retinal Vessel Segmentation via Locally Adaptive Derivative Frames in Orientation Scores". In: *IEEE Transactions on Medical Imaging* 35.12 (Dec. 2016), pp. 2631–2644. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2016.2587062. URL: http://ieeexplore.ieee.org/document/7530915/ (visited on 07/20/2021).

[86] Zhuo Zhang et al. *A survey on computer aided diagnosis for ocular diseases*. Aug. 31, 2014. DOI: 10.1186/1472-6947-14-80. URL: https://doi.org/10.1186/1472-6947-14-80 (visited on 06/28/2021).

[87] Rongchang Zhao et al. "Multi-index Optic Disc Quantification via MultiTask Ensemble Learning". In: *MICCAI*. 2019.

[88] Rongchang Zhao et al. "Weakly-Supervised Simultaneous Evidence Identification and Segmentation for Automated Glaucoma Diagnosis". In: *AAAI*. 2019.