Titre: Title:	Dépistage automatique de la rétinopathie diabétique dans les images de fond d'oeil à l'aide de l'apprentissage profond
Auteur: Author:	Youri Boris Peskine
Date:	2020
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Peskine, Y. B. (2020). Dépistage automatique de la rétinopathie diabétique dans les images de fond d'oeil à l'aide de l'apprentissage profond [Master's thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/5415/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/5415/
Directeurs de recherche: Advisors:	Farida Cheriet
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Dépistage automatique de la rétinopathie diabétique dans les images de fond d'œil à l'aide de l'apprentissage profond

YOURI BORIS PESKINE
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Août 2020

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Dépistage automatique de la rétinopathie diabétique dans les images de fond d'œil à l'aide de l'apprentissage profond

présenté par **Youri Boris PESKINE**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Thomas HURTUT, président
Farida CHERIET, membre et directrice de recherche
Renaud DUVAL, membre externe

DÉDICACE

à ma famille et mes amis, sans qui tout ceci aurait été impossible

REMERCIEMENTS

Je tiens à remercier ma directrice de recherche, Professeure Farida Cheriet, pour cette opportunité de projet. J'ai énormément appris lors de ce travail, que ce soit sur le plan professionnel ou social, j'ai pu découvrir le monde passionnant qu'est la recherche en informatique sous sa direction.

Je remercie également le CIHR SPOR Network in Diabetes and its Related Complications (DAC) pour le support financier.

Je souhaite aussi remercier tous les membres du laboratoire LIV4D pour les aides précieuses et les conseils éclairés qu'ils m'ont apportés. Chacun de leur commentaires pertinents m'ont permis de grandement progresser.

Enfin, je tiens à remercier ma famille pour m'avoir supporté pendant toute la durée de mes études, et surtout pendant ces deux années loin d'eux. Ce soutient régulier m'a permis de me dépasser au quotidien et de m'épanouir.

RÉSUMÉ

Le diabète est une maladie chronique qui touche plus de 400 millions d'adultes dans le monde. Cette maladie peut entraîner plusieurs complications au cours de la vie d'un malade. Une de ces complications est la rétinopathie diabétique. Il s'agit de la principale cause de cécité chez l'adulte. Cette maladie apparaît souvent sans symptômes, il est donc important pour les personnes atteintes de diabète d'effectuer des vérifications régulières chez un ophtalmologue. Cette vérification s'effectue par la prise d'images numériques de fond d'oeil du patient. Ces images sont ensuite examinées par un médecin afin de donner un diagnostic.

Dans ce travail, il est question d'automatiser le diagnostic de la rétinopathie diabétique à l'aide des images numériques de fond d'oeil ainsi que l'apprentissage profond. En effet, les réseaux de neurones ont suscité ces dernières années un intérêt important dans différents domaines, notamment ceux du médical et de la vision par ordinateur. Les réseaux convolutifs permettent des applications tel que la classification ou la segmentation d'images. Ici, la classification correspond à classer les images selon la présence ou non de la rétinopathie diabétique dans les images et la segmentation correspond à extraire des régions d'intérêt, comme les vaisseaux sanguins par exemple. Cependant, certaines images ne sont pas de qualité suffisante pour effectuer un diagnostic. Le Scottish Diabetic Retinopathy Grading Scheme (SDRGS) décrit les normes de qualité des images de fond d'oeil. Selon ce document, la qualité des images est importante puisqu'une image de mauvaise qualité ne peut pas être diagnostiquée. L'évaluation de la qualité des images de fond d'oeil est donc tout aussi important que l'automatisation de leur diagnostic.

Nous pouvons ainsi définir comme objectif global celui d'automatiser le diagnostic de la rétinopathie diabétique dans les images de fond d'oeil. Cela implique aussi l'évaluation de la qualité de ces images de fond d'oeil.

L'évaluation de la qualité des images est basée sur l'extraction de régions d'intérêt dans les images de fond d'oeil, la macula ainsi que les vaisseaux présents autour de la fovéa. Ces régions seront extraites à l'aide de réseaux convolutifs effectuant de la segmentation. Cette méthode d'évaluation établit un score de qualité pour une image. Il est important que ce score soit interprétable afin de mieux comprendre les résultats. Cette approche est beaucoup inspirée de celle utilisée par les cliniciens pour évaluer la qualité des images. Cet algorithme a obtenu 100% de sensibilité et 93% de spécificité sur une base de données de 88 images.

Ensuite, le diagnostic des images est effectué avec des réseaux de neurones convolutifs effectuant de la classification et de la régression. Ces réseaux sont entraînés sur des ensembles

d'images de fond d'oeil, notamment lors d'une compétition Kaggle. Le modèle en question a obtenu un kappa de 0.915 sur la base de données de test privé de cette compétition. Nous avons ensuite étudié l'impact de la qualité des images lors des phases d'entraînement et de test des modèles. Ceci nous a permis de remarquer que la qualité des images a beaucoup d'importance lors de la phase de test.

Ce travail a pour but d'aider le processus de dépistage de la RD en ajoutant des outils d'automatisation. La méthode permettant d'évaluer la qualité des images numériques peut être un outil intéressant lors du processus d'acquisition des images afin de ne plus transmettre aux cliniciens des images de mauvaise qualité. En effet, cette méthode permettrait aux techniciens de s'assurer que les images prises sont de bonne qualité et de reprendre certaines images lorsqu'il est nécessaire. Le module de dépistage automatique peut permettre un processus plus facile dans les zones où le dépistage est difficile pour obtenir un aperçu rapide de la situation.

ABSTRACT

Diabetes is a chronic disease that currently concerns more than 400 millions of adults in the world. This disease can cause several complications during the life of a person. One of these complications is diabetic retinopathy. Being one of the leading cause of blindness in the working age population, this complication is serious and requires medical prevention. This disease often appear without any symptoms, meaning that regular examinations with an ophtalmologist are required to enable its detection and treatment.

This work is about automating the diagnostic of diabetic retinopathy, with the use of digital fundus images and deep learning. Indeed, deep learning and neural networks have recently been used in several fields, such as medical applications or computer vision. Convolutional neural networks perform applications such as image classification or image segmentation really well. Here, classification means to label each image based on the presence or absence of diabetic retinopthy in the images and segmentation means to extract regions of interest in the image, such as blood vessels. However, some images do not meet the quality requirements to be diagnosed. The Scottish Diabetic Retinopthy Grading Scheme describes the quality norms in fundus images. According to this document, evaluation of fundus image quality is mandatory, because a bad quality image cannot be diagnosed. Evaluation of the image quality is as important as the automation of the screening.

This work can therefore be defined with two main goals : the evaluation of fundus image quality and the automatic screening.

The evaluation of the fundus image quality is based on segmenting regions of interest in those images. These regions are the macula as welle as the small vessels radiating around the fovea. Segmentation models are used to extract these regions. Once these regions are extracted, a score is computed to match the overall quality of the image. This score needs to be as much interpretable as possible, as we need to understand why an image is good or bad quality. This approach is deeply inspired by the way clinicians assess fundus image quality. The method achieved 100% sensitivity and 93% specificity on a dataset containing 88 images.

Then, the grading of the fundus images is done with convolutional neural networks. The objectives of these neural networks are classification or regression. These networks are trained using different dataset and during a Kaggle challenge. Our model achieved a kappa of 0.915 on the private test dataset of this challenge. We then studied the impact of image quality during the training and testing of our models. We noticed that image quality has a significant

role during testing.

This work aims to help the screening procedure of diabetic retinopathy by introducing some automatisation tools. The quality evaluation method could be used to improve the image aquisition process. This work can help technicians better assess the image quality and retake images when necessary. The screening model could be used to allow an easier diagnosis in certains difficult areas to quickly obtain an overview of the situation.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

RD	Rétinopathie Diabétique
SDRGS	Scottish Diabetic Retinopathy Grading Scheme
CNN	Convolutional Neural Networks (réseaux de neurones convolutifs)
MLP	Multi Layer Perceptron (perceptron multicouche)
SVM	Support Vector Machine (machine à vecteur de support)
TTA	Test-time augmentation (augmentation de données lors de la phase de test)

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

1.1 Contexte

Selon l'organisation mondiale de la santé, environ 422 millions de personnes sont atteintes de diabète en 2017 [4], et la prévalence de cette maladie est en forte augmentation, notamment dans les pays pauvres et en voie de développement. Le diabète est une maladie chronique liée au manque d'insuline. Il s'agit d'une hormone sécrétée par le pancréas et son rôle est de favoriser l'absorption du glucose par certaines cellules. Un manque d'insuline signifie que le glucose ne peut pas être absorbé par ces cellules. Ceci se traduit par un taux élevé de sucre dans le sang et entraîne de nombreuses complications. Parmi ces complications, la fragilisation des parois des vaisseaux sanguins de la rétine est celle qui cause la RD. La rétine est une membrane fine sensible à la lumière au fond du globe oculaire qui transforme les signaux lumineux en influx nerveux à l'aide de photorécepteurs. Chez une personne atteinte de RD, les vaisseaux rétiniens sont fragiles et entraînent des lésions. Ceci cause une perte progressive de la vision, surtout lorsque la zone de la macula est atteinte. La macula est une zone du fond de l'oeil, proche du centre de la rétine. Cette zone est responsable de notre vision centrale et son bon fonctionnement est très important. La figure 1.1 présente l'anatomie de l'oeil. La RD entraîne la baisse de l'acuité visuelle jusqu'à la cécité dans les cas les plus graves. Il s'agit d'une des principales causes de perte de vue et touche 25-30% des gens atteints de diabètes, soit environ 120 millions de personnes dans le monde. Heureusement, des traitements efficaces existent afin de ralentir la maladie.

La RD peut apparaître sans aucun symptôme, c'est pourquoi il est important d'effectuer des examens régulier en prévention pour contrôler son avancée. Lors de ces examens, des images numériques du fond d'oeil du patient sont prises à l'aide de caméras spécialisées. En général, il est recommandé aux techniciens de prendre trois images par oeil : deux centrées sur la macula, et une centrée sur la papille. Ces images sont ensuite examinées par un clinicien permettant de donner un diagnostic sur l'avancée de la RD. Il existe plusieurs méthodes pour diagnostiquer la RD. Au Quebec, le Scottish Diabetic Retinopathy Grading Scheme (SDRGS) [5] est utilisé. Ce document explique en détail comment diagnostiquer la RD. Dans ce travail, nous suivons en majeur partie les règles proposées par le SDRGS.

Ce document explique en détail quels sont les différents niveaux d'avancée de la maladie et les différents critères sur la qualité. Ces informations sont résumées dans le tableau 1.1. La sévérité de la maladie est classée sur 5 niveaux : R0 (absent), R1 (légère), R2 (modérée), R3 (non proliférante sévère) et R4 (proliférante). Ces niveaux dépendent de caractéristiques

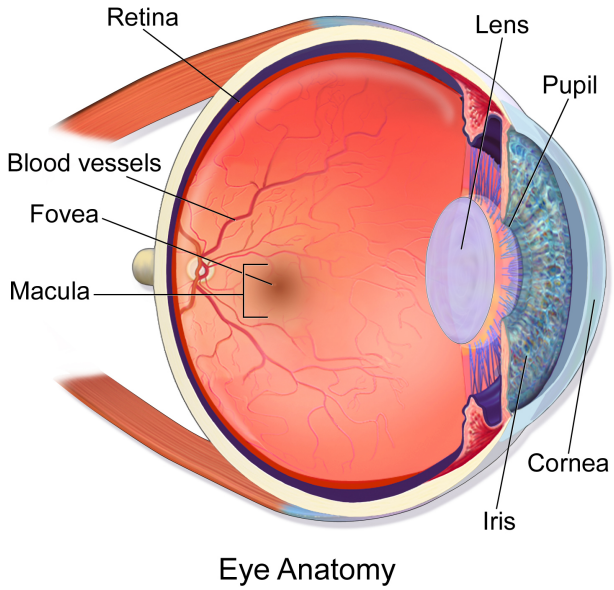


Figure 1.1 Schéma de l'oeil [1]

comme des hémorragies, des exudats ou des microanévrysmes. La figure 1.2 présente ces caractéristiques. Le niveau R5 correspond aux personnes ayant un oeil énuclée, il ne sera donc pas important pour notre travail. Le niveau R6 correspond aux cas où aucune des images numériques n'est de qualité suffisante pour donner un diagnostic.

Les critères de qualité sont aussi définis par le SDRGS et sont les suivants :

1. La région photographiée correspond à la bonne région de la rétine :
   — L'entièreté du disque optique doit être visible.
   — La fovea doit être éloignée d'une valeur supérieure à 2 diamètres du disque optique des bords de l'image. (Figure 1.3a)
2. La clarté de l'image est adéquate :
   — La troisième génération de vaisseaux présents autour de la fovéa doit être visible. (Figure 1.3b)

La figure 1.3 présente ces deux critères de qualité.

L'automatisation du diagnostic de la RD est un problème important, permettant d'aider les cliniciens à diagnostiquer les patients. L'avancée des réseaux de neurones et de l'apprentissage profond ces dernières années a permis d'obtenir des résultats comparables à ceux des cliniciens. De plus, le dépistage automatique permet un diagnostic plus accessible dans les pays pauvres et en voie de développement.

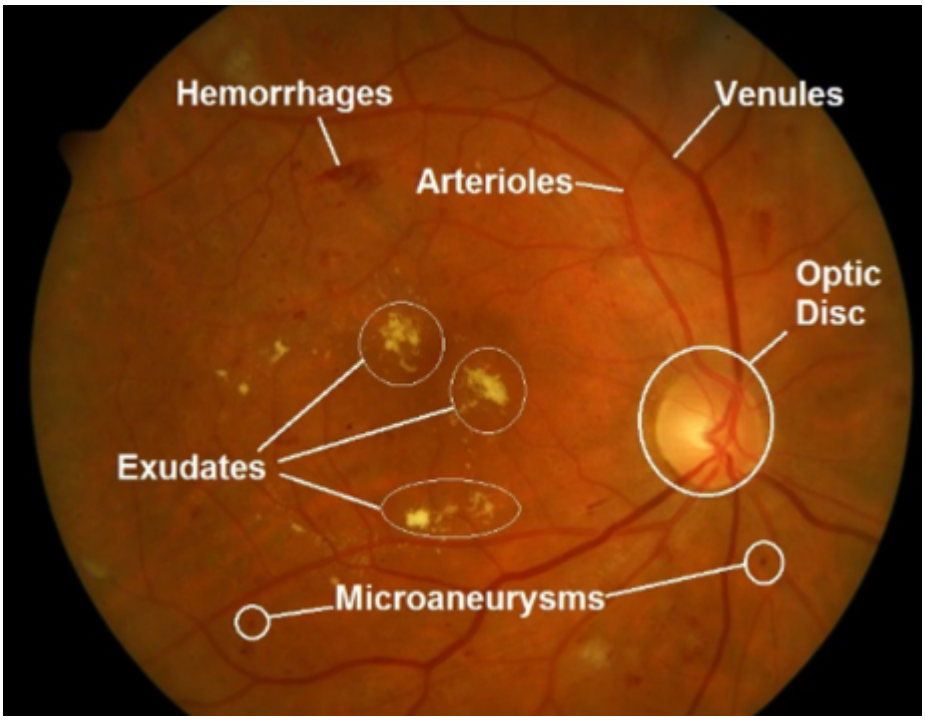


Figure 1.2 Caractéristiques qui définissent le niveau de sévérité de la RD par Abdullah et al [2].

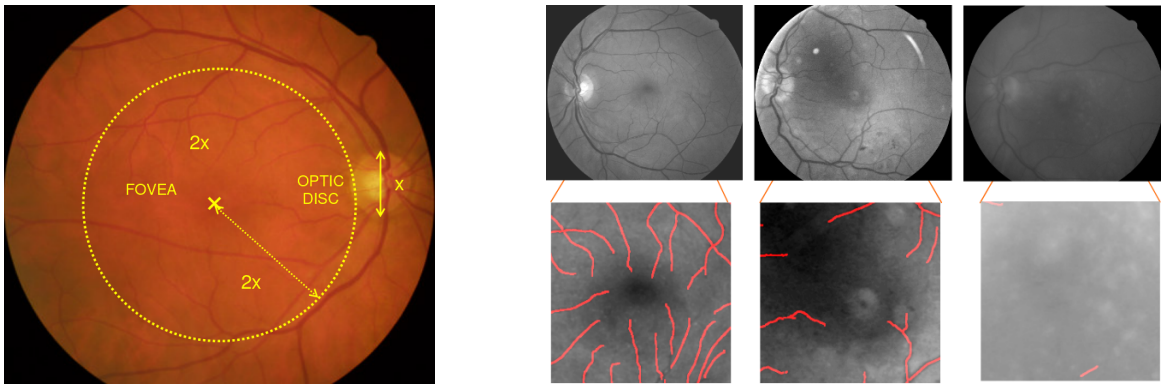
1.2 Objectifs de recherche

L'objectif global de ce mémoire est d'automatiser le diagnostic de la RD dans les images de fond d'oeil. Pour ce faire, nous devons réaliser plusieurs sous-objectifs.

En premier lieu, il est question d'établir un modèle évaluant la qualité des images de fond d'oeil. Les images de mauvaise qualité sont une perte de temps et de ressources pour les cliniciens, les techniciens et les patients. Ce modèle nous permettra d'évaluer la qualité d'une image en lui associant un score de qualité. Il doit pouvoir être interprétable pour nous permettre de comprendre les résultats. Il devra donc suivre les règles imposées par le SDRGS.

Ensuite, nous concevrons un modèle permettant de diagnostiquer ces images en fonction de leur niveau de maladie. Nous étudierons les performances de différents types de réseaux de neurones, ainsi que plusieurs stratégies d'apprentissage. Ce modèle permet de prédire le niveau d'avancée de la maladie dans une image.

Enfin, il est question d'étudier l'impact de la qualité des images dans les phases d'apprentissage et d'entraînement des modèles d'apprentissage profond.



(a) (b)

Figure 1.3 Critères de qualité définis par le SDRGS. (a) La fovea doit être éloignée d'au moins 2 diamètres du disque optique des bords de l'image. (b) La troisième génération de vaisseaux présents autour de la fovéa doit être visible. Figures attribuées à [3]

1.3 Plan du mémoire

Le prochain chapitre présente d'abord les concepts importants de l'apprentissage machine. Ensuite, nous aborderons les travaux étudiés dans la littérature sur l'évaluation de la qualité des images de fond d'oeil et dans le diagnostic de la RD. Nous détaillerons les méthodes retenues pour effectuer ces objectifs dans le chapitre 3. Nous discuterons ensuite des résultats des méthodes dans le chapitre 4. Enfin, nous présenterons les limitations et les les travaux futurs dans une conclusion.

Tableau 1.1 Niveaux de la rétinopathie diabétique définis dans le Scottish Diabetic Retinopathy Grading Scheme

Niveau	Description	Résultats
R0 (absent)	Pas de signe de la maladie	Refaire le test de dépistage dans 12 mois
R1 (légère)	Présence d'au moins une des caractéristiques suivantes : points hémorragiques, micro-anévrismes, exsudats durs, nodules cotonneux, hémorragies en taches, hémorragies superficielles / en flammèches	Refaire le test de dépistage dans 12 mois
R2 (modérée)	4 ou plus hémorragies en taches présentes dans un hémichamp* seulement (supérieur ou inférieur)	Refaire le test de dépistage dans 6 mois
R3 (non proliférante sévère)	Présence d'au moins une des caractéristiques suivantes : 4 ou plus hémorragies en taches dans chaque hémichamp (inférieur et supérieur), chapelets veineux, anomalies microvasculaires intrarétiniennes (AMIR)	Orienter vers un spécialiste
R4 (proliférante)	Présence d'au moins une des caractéristiques suivantes : néovaisseaux au niveau de la papille optique, néovaisseaux ailleurs, hémorragie intravitréenne, décollement de rétine	Orienter vers un spécialiste
R5 (énuclée)	Oeil énuclée	Refaire le test de dépistage dans 12 mois (autre oeil)
R6 (inadéquat)	Qualité insuffisante : La rétine n'est pas suffisamment visible pour l'évaluation de la RD	Echec technique. Organiser un examen de dépistage alternatif

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans ce chapitre, nous allons détailler les notions importantes et les études réalisées dans le domaine. Tout d'abord, nous allons introduire certaines notions importantes de l'apprentissage machine et donner leur définitions. Ensuite, nous aborderons le problème d'évaluation de la qualité des images de fond d'oeil. Enfin, nous présenterons les méthodes de dépistage automatique.

2.1 L'apprentissage machine

L'apprentissage machine est le domaine informatique regroupant les algorithmes pouvant s'adapter à un problème en apprenant d'exemples donnés. Ces algorithmes ou modèles apprennent de représentations ou de motifs dans des données afin de leur permettre de résoudre certaines fonctions. Cet apprentissage peut être supervisé ou non, selon les données à dispositions et la tâche à effectuer. Dans notre travail, nous abordons l'apprentissage supervisé pour nos modèles.

2.1.1 L'apprentissage supervisé

Il s'agit d'un apprentissage basé sur des exemples annotés, formant des paires entrée - sortie qui constituent un ensemble d'entraînement pour le modèle. Pour chaque exemple, le modèle va essayer de prédire la sortie sachant l'entrée. Ce modèle peut donc être représenté par une fonction $f : x \rightarrow \hat{y}$ qui prédit la sortie $\hat{y}$ sachant l'entrée x. De plus, on peut aussi définir une fonction de perte $\mathcal{L} : (\hat{y}, y) \rightarrow l$ qui permet de quantifier l'erreur l entre la prédiction de notre modèle $\hat{y}$ et la véritable sortie y correspondant à l'entrée x. Dans le cadre de l'apprentissage supervisé, le but du modèle est de minimiser cette fonction $\mathcal{L}$ pour tout l'ensemble d'entraînement. Il existe de nombreuses fonctions de pertes selon le problème à résoudre et les fonctions utilisés pour l'apprentissage de nos modèles seront présentées dans la section 2.1.4.

2.1.2 Les réseaux de neurones et l'apprentissage profond

Les réseaux de neurones sont un type de modèles très répandus pour effectuer ce genre de tâche. Originalement inspirés des neurones biologiques du cerveau, les neurones artificiels sont des fonctions mathématiques simples. Les connexions entre les neurones, inspirées des synapses biologiques, sont des poids qui multiplient les sorties des neurones. Un réseau est

structuré en couches ordonnées de neurones. Les neurones d'une couche sont connectés aux neurones des couches précédente et suivante. Les données en entrée parcourent donc le réseau de couches en couches jusqu'à atteindre la couche de sortie qui correspond au résultat. Lorsque le nombre de couches est important, on parle d'apprentissage profond. Le perceptron multicouche (MLP) est un type de réseau de neurone où tous les neurones de couches adjacentes sont connectés entre eux. Ces couches sont dites « complètement connectées » . La figure 2.1 présente ce type d'architecture.

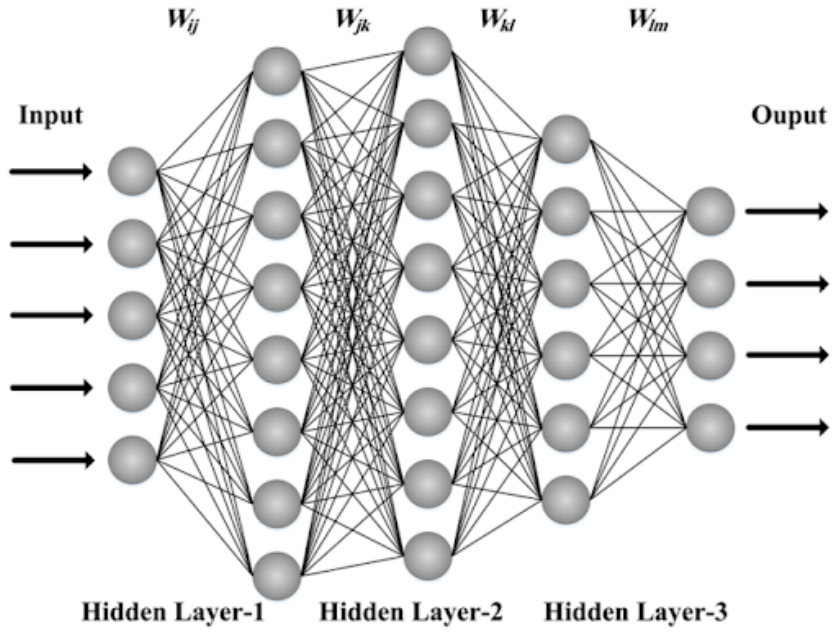


Figure 2.1 Schéma d'un perceptron multicouche. Le vecteur d'entrée est propagé au sein du réseau, Chaque synapse correspond à une multiplication et chaque neurone possède une fonction d'activation.

L'apprentissage est effectué à l'aide d'algorithmes de rétro-propagation. Ces algorithmes calculent la perte l entre la prédiction du réseau $\hat{y}$ et la solution y à l'aide de la fonction $\mathcal{L}$ pour un exemple donné (x, y). Ensuite, les paramètres de chaque neurone est ajusté afin de minimiser $\mathcal{L}$ et donc de rapprocher $\hat{y}$ de la solution y. Cette opération est effectuée plusieurs fois pour tous les exemples (x, y) d'entraînement.

2.1.3 Les réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN) sont des réseaux très largement utilisés en apprentissage machine, et surtout lorsque les données sont des images. Ce type de réseau se distingue du MLP par des couches de convolution. Les images sont des données de grande

taille, ce qui augmente énormément le nombre de connexions entre les couches d'un MLP. Les couches de convolution permettent de réduire ce nombre de connexions trop élevé tout en conservant les informations importantes. Ces couches utilisent des opérateurs convolutifs, semblables à des filtres, qui conservent les corrélations locales dans l'image. Il s'agit d'opérateurs très largement utilisés dans le domaine de traitement des signaux, et donc appropriés aux images.

Ces réseaux de neurones sont utilisés dans le domaine de la vision par ordinateur pour effectuer différentes taches, comme la classification et la segmentation. La classification a pour but de séparer les images en un certain nombre de classes. Dans notre cas, il s'agira de répartir les images selon le tableau 1.1, ou plus simplement de distinguer un patient malade d'un patient sain. La segmentation consiste à extraire certaines régions dans les images. En d'autre termes, il s'agit de classifier chacun des pixels de l'image. Cela permet d'extraire des régions d'intérêt dans une image, comme les vaisseaux sanguins par exemple.

La performance des CNN sur les tâches de vision par ordinateur a révolutionné les méthodes en classification et en segmentation d'images. Krizhevsky et al. [6] montrent en 2012 la capacité des CNN larges à effectuer de la classification d'images sur la base de données ImageNet [7] en utilisant l'apprentissage supervisé. Cette base de données contient plus de 15 millions d'images naturelles annotées réparties en plus de 22000 classes, il s'agit de la référence en terme de classification d'image. En général, 1000 classes sont utilisées, avec environ 1000 images par classe. De nombreux autres travaux ont ensuite amélioré les performances des CNN pour la classification d'images sur ImageNet en introduisant des architecture différentes, comme les réseaux AlexNet [6], VGG [8], Inception [9], ResNet [10] ou EfficientNet [11]. Chaque architecture est différente et introduit de nouvelles opérations. Dans notre travail, il sera question des réseaux ResNet et EfficientNet en majeure partie.

Le réseau ResNet a pour but de simplifier l'apprentissage, en introduisant des fonctions résiduelles, simplifiant l'optimisation. En effet, l'optimisation de réseaux très larges est complexe, et augmenter le nombre de couches ne suffit pas pour augmenter les performances d'un réseau. Le concept de ces fonctions résiduelles est d'introduire des connections raccourcies qui sautent des couches de convolutions. La figure 2.2 présente l'architecture d'un ResNet-34, constitué d'un faible nombre de paramètres à optimiser comparé a un réseau VGG, tout en obtenant de très bonnes performances. Le chiffre 34 correspond au nombre de couches dans le réseau. Il existe d'autres ResNets avec un nombre de couches différentes, comme ResNet-18, ResNet-50 ou ResNet-101 par exemple.

D'avantages d'ajouts ont été effectués sur cette architecture. Par exemple, l'ajout de couches résiduelles agrégées donnent naissance aux ResNeXt [12] et l'ajout de couche de compression

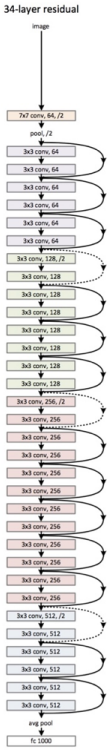


Figure 2.2 L'architecture d'un réseau ResNet-34

et d'excitation donne lieu au modèle SE-ResNeXt [13].

Les réseaux EfficientNet cherchent aussi à optimiser l'apprentissage des réseaux convolutifs. L'idée de ce travail est de changer la manière dont les réseaux sont agrandis. En général, il existe trois approches pour augmenter la taille d'un réseau convolutif : augmenter le nombre de couches (profondeur), augmenter la taille des couches (épaisseur), ou augmenter la résolution des images d'entrée. Dans ce travail, il est montré qu'il existe un équilibre entre ces trois paramètres afin d'optimiser l'agrandissement d'un réseau. Aussi, une architecture de base adaptée à ce type d'augmentation appelée EfficientNet est développée. Le modèle de base EfficientNet-B0 est agrandi en épaisseur, en profondeur et en résolution pour donner les réseaux EfficientNet-B1 à EfficientNet-B7.

Tout comme la classification d'images, le domaine de la segmentation d'images utilise aussi des CNN mais les architectures sont différentes. Le U-Net [14] publié en 2015 est un réseau qui révolutionne la segmentation d'images, notamment dans le domaine médical. Son architecture est basée sur deux parties, une première permettant de capturer le contexte et une seconde permettant de capturer la localisation. Ce type de réseau est très performant et ne requiert pas un nombre très important d'images d'entraînement pour fonctionner, ce qui en fait un modèle très fiable.

2.1.4 Métriques

Plusieurs métriques sont utilisées afin de calculer les performances d'un modèle informatique. Les plus importantes sont le kappa de Cohen [15], la sensibilité, la spécificité, la précision, le rappel, et les fonctions de perte $\mathcal{L}$.

Kappa de Cohen

Le kappa de Cohen permet de mesurer un accord entre deux listes de diagnostics. Cette métrique permet de mesurer si les diagnostics prédits par notre modèle correspondent aux diagnostics des cliniciens. Cette métrique est calculée à l'aide de la formule suivante :

$$\kappa = 1 - \frac{\sum_i \sum_j w_{ij} c_{ij}}{\sum_i \sum_j w_{ij} m_{ij}} \tag{2.1}$$

avec W la matrice des poids, C la matrice de confusion entre les listes et M la matrice du produit cartésien entre les histogrammes des deux listes. En général, les poids utilisés sont quadratiques, ce qui signifie que les erreurs de diagnostic entre deux classes proches ont un léger impact sur le kappa, et les erreurs entre deux classes éloignées ont un impact plus important.

Les valeurs possibles du kappa vont de -1 à 1. Un kappa de 1 signifie un accord parfait entre les listes, un kappa de -1 signifie un accord opposé entre les listes. Un kappa de 0 signifie que l'accord entre les listes est absent, ici cela signifie que le modèle n'est pas plus performant que le hasard.

Sensibilité et spécificité

La sensibilité et la spécificité sont deux métriques qui permettent de mesurer les performances d'un modèle lors d'une classification binaire, comme le tri de la qualité des images. Ces métriques utilisent la matrice de confusion binaire C entre les listes pour le calcul. Cette matrice contient le taux de faux positif (fp), faux négatif (fn), vrai positif (vp) et vrai négatif (vn).

La sensibilité mesure la proportion d'exemples positifs qui ont été correctement identifiés. L'équation est la suivante :

$$se = \frac{vp}{vp + fn} \tag{2.2}$$

A l'inverse, la spécificité mesure la proportion d'exemples négatifs qui ont été correctement identifiés. L'équation est la suivante :

$$sp = \frac{vn}{vn + fp} \tag{2.3}$$

Précision et rappel

La précision et le rappel sont des mesures complémentaires à la sensibilité et à la spécificité. Elles mesurent aussi les performances d'un modèle lors d'une classification binaire et utilisent la matrice de confusion C.

La précision mesure la proportion d'exemples réellement positifs parmi les exemples détectés positifs. L'équation est la suivante :

$$p = \frac{vp}{vp + fp} \tag{2.4}$$

Le rappel mesure la proportion d'exemples positifs détectés. L'équation est la suivante :

$$r = \frac{vp}{vp + fn} \tag{2.5}$$

Les fonctions de pertes $\mathcal{L}$

Pour un problème de classification d'images, la fonction de perte $\mathcal{L}_{class}$ la plus couramment utilisée est l'entropie croisée. Elle permet des prédictions discrètes et est définie selon la formule suivante :

$$\mathcal{L}_{class}(x, c) = -\log \frac{exp(x[c])}{\sum_j exp(x[j])} \tag{2.6}$$

avec x un vecteur contenant le score de chaque classe et c l'index de la classe désirée.

Nous pouvons aussi utiliser un modèle effectuant des prédictions continues, car les classes en question sont ordonnées. Pour ce faire, nous utiliserons un modèle ayant pour objectif la régression. La fonction de perte $\mathcal{L}_{reg}$ pour un modèle de régression est l'erreur quadratique moyenne. Elle est définie selon la formule suivante :

$$\mathcal{L}_{reg}(x, c) = mean[(x_1 - c_1)^2, ..., (x_n - c_n)^2] \tag{2.7}$$

avec x un vecteur contenant les prédictions de n exemples et c un vecteur contenant les n classes désirées.

2.1.5 Pré-traitement

Le pré-traitement est un moyen de corriger certains défauts des images pour un meilleur apprentissage. En général cela se traduit par l'application de filtres d'images ou des opérations sur l'histogramme de l'image. Cela permet aux images d'être semblables et agit comme une normalisation en entrée du réseau.

2.1.6 Augmentations de données

Un autre moyen d'augmenter les performances d'un modèle et d'effectuer de l'augmentation de données. L'augmentation de données est un processus permettant aux modèles de mieux généraliser. En pratique, il s'agit d'effectuer des transformations sur une image lors de l'entraînement, comme des rotations, des agrandissements d'image ou des retournements. Ceci permet d'augmenter la variabilité d'une base de données à partir des images d'origine.

2.1.7 Méthodes ensemblistes

Différentes méthodes ensemblistes permettent d'augmenter d'avantage les performances, après l'entraînement des modèles. La première méthode ensembliste utilisée est l'augmentation de données lors de la phase de test (test time augmentation ou TTA en anglais). Cette méthode consiste à appliquer les transformations d'images utilisées pour l'augmentation de données lors de la phase d'entraînement sur les images de test. Ces transformations sont effectuées sur plusieurs copies d'une même image de test et le modèle effectue les prédictions sur chacune des images. Ceci forme un ensemble de prédictions pour une même image de test. Cet ensemble est ensuite utilisé pour la prédiction finale de l'image. Si le modèle a un objectif de classification, les prédictions sont discrètes et la prédiction finale est obtenue par un système de vote : la prédiction obtenant la majorité est la prédiction finale. Si le modèle a un objectif de régression, les prédictions sont continues et la prédiction finale peut être obtenue en arrondissant la moyenne des prédictions.

Une autre méthode ensembliste consiste à regrouper les prédictions de plusieurs modèles différents. L'arrondi de la moyenne des prédictions de chaque modèle pour une image de test correspond à la prédiction finale de l'image. Cette méthode utilise la diversité de différents modèles et regroupe des prédictions avec des variations. En général, on cherche à regrouper des modèles différents pour qu'ils soient complémentaires.

2.2 La qualité des images de fond d'oeil

Une revue de littérature complète concernant les algorithmes d'évaluation de la qualité des images de fond d'oeil a été faite par A.Raj et al. [16] en 2019. Elle détaille de manière exhaustive les principales méthodes permettant l'évaluation de la qualité des images de fond d'oeil. Cette revue divise ces algorithmes en trois principales catégories : basées sur la similarité, sur la segmentation et sur l'apprentissage machine. Nous allons étudier chacune de ses catégories afin d'en comprendre les avantages et les inconvénients.

2.2.1 Méthodes basés sur la similarité

Les méthodes basées sur la similarité comparent des caractéristiques des images de fond d'oeil avec celles d'une base d'images préétabli comme étant de bonne qualité. Il s'agit de la première approche d'évaluation de la qualité des images de rétine, par Lee et Wang [17]. Dans ce travail, l'histogramme de l'image à évaluer est comparé avec les histogrammes d'images de bonne qualité en utilisant le produit de convolution. L'avantage de cette technique est sa simplicité d'implémentation. Cependant, elle est difficilement généralisable car la création d'une base complète d'images de bonne qualité est complexe. Aussi, cette approche basée sur les histogrammes perd toute information locale dans l'image, et n'utilise pas les structures importantes. Cette approche ne se base pas sur les connaissances de l'anatomie, elle n'a pas de réelle interprétation clinique.

Lalonde et al. [18] proposent une méthode basée sur la similarité utilisant la distribution de l'amplitude des contours ainsi que la distribution d'intensité locale. Contrairement à [17], cette approche conserve certaines informations locales. Cependant, elle souffre des mêmes problèmes de généralisation. Aussi, les structures étudiées ici ne sont pas celles étudiées dans le SDRGS, ce qui rend l'interprétation de la méthode difficile.

2.2.2 Méthodes basées sur la segmentation

Les méthodes basées sur la segmentation vont extraire des régions d'intérêt dans les images afin de déterminer la qualité de l'image. Ces méthodes sont généralement séparées en deux parties : la segmentation et l'analyse. Par exemple, A.Hunter et al. [19] segmentent les vaisseaux sanguins présents dans l'image. Ensuite, ils utilisent le contraste de ces vaisseaux afin d'évaluer la qualité des images. L'approche de Fleming et al. [20] est basée sur l'analyse des petits vaisseaux présents autour de la fovea. La localisation de la macula et la segmentation des vaisseaux est effectuée à l'aide d'outils classiques du traitement d'images comme l'utilisation de la transformée de Hough et l'analyse du contraste. L'avantage de ces méthodes basées

sur la segmentation est leur réel interprétation clinique. En effet, ces méthodes se rapprochent le plus du SDRGS. Cependant, la plupart de ces méthodes utilisent des outils classiques du traitement d'images comme les filtres, les transformées, les analyses de contraste ou les histogrammes. En général, ces fonctions utilisent des paramètres ou des variables qui dépendent d'une base de données en particulier. Il est difficile pour ces méthodes de généraliser pour prendre en compte une plus forte variabilité des données sans changer les paramètres et les variables.

2.2.3 Méthodes basées sur l'apprentissage machine

Les méthodes basées sur l'apprentissage machine peuvent encore être séparées en deux sous-catégories : les algorithmes basés sur des caractéristiques sélectionnées en amont et ceux basés sur des caractéristiques de l'apprentissage profond. Par exemple, dans [21–23] les caractéristiques sélectionnées en amont sont la couleur, la texture ou la netteté. Ces caractéristiques ne correspondent pas à la définition de la qualité du SDRGS, il s'agit seulement de critères de qualité pour des images naturelles. L'apprentissage profond est utilisé par [24, 25] pour extraire des caractéristiques et ainsi évaluer la qualité des images. Cependant, les caractéristiques extraites par l'apprentissage machine sont souvent critiquées pour être des « boites noires » qui sont peu interprétables. L'avantage des méthodes basées sur l'apprentissage machine est la généralisation puisqu'elles utilisent souvent des bases de données volumineuses.

2.3 L'utilisation de l'apprentissage machine pour le diagnostic de la rétinopathie diabétique

2.3.1 Le diagnostic automatique

Différentes méthodes d'automatisation du diagnostic de la RD ont été étudiées. Les premières approches utilisaient des outils classiques d'apprentissage machine comme les machines à vecteur de support (SVM). Acharya et al. [26] utilisent un SVM avec des caractéristiques de texture pour effectuer la classification de la DR. Noronha et al. [27] utilisent un SVM ainsi que les caractéristiques d'une transformation d'ondelette. Ces méthodes sont graduellement remplacées par l'utilisation d'apprentissage profond, comme les CNN par exemple.

L'utilisation de CNN afin de diagnostiquer des maladies à partir d'images médicales se répand progressivement dans différents domaines comme celui de la radiologie [28], la dermatologie [29] ou l'ophtalmologie. Le diagnostic automatique de la RD a notamment été abordé lors d'une compétition web en 2015 sur le site Kaggle [30]. Les données de cette compétition consistent en une base de données d'entraînement et une base de données de test. La base

de données d'entraînement regroupe plus de 30000 images réparties en 5 niveaux de RD. Le base de données de test contient plus de 50000 images. La métrique utilisée pour évaluer les performances d'un modèle est le kappa de Cohen, décrit par l'équation 2.1. Cette métrique permet de mesurer l'accord entre les valeurs réelles et celles prédites par les modèles. Un kappa de 1 signifie un accord parfait et un kappa de 0 signifie une absence d'accord. Lors de la compétition Kaggle, le meilleur modèle a obtenu un kappa de 0,850. En 2019, une nouvelle compétition Kaggle [31] a été organisée à laquelle nous avons participé.

Le gagnant de la compétition Kaggle 2019 détaille son approche sur le site [32]. Dans ce document, il dit fusionner les base de données Kaggle 2015 et 2019 pour l'entraînement de ses modèles. Il est important de noter qu'il n'applique aucun pré-traitement aux images mais il effectue un grand nombre d'augmentations de données. Le modèle est un ensemble de modèles Inception et SE-ResNeXt. L'auteur aurait aimé ajouter des réseaux EfficientNet s'il en avait le temps.

Les méthodes de diagnostic automatique se basant sur les CNN progressent en même temps que ces derniers. La base de données publique de Kaggle 2015 est souvent utilisé par les travaux sur le diagnostic automatique de la RD. Par exemple, Lam et al. [33] utilisent les réseaux Inception [9] et AlexNet [6] préentraînés sur ImageNet et continuent l'entraînement sur la base de données Kaggle 2015 pour effectuer un diagnostic de RD. Dans ce travail, le modèle est entraîné plusieurs fois pour effectuer des classifications à 2, 3 et 4 classes. Il est notable que les performances décroissent avec le nombre de classes, car la tâche de discerner deux classes devient plus difficile car les détails sont plus subtils. De plus, il est aussi abordé l'importance de la base de données d'entraînement ainsi que du pré-traitement sur les résultats des modèles. Le réseau VGG [8] est aussi utilisé sur la base de données Kaggle 2015 par Rakhlin et al. [34] pour effectuer le diagnostic de la RD. Dans ce travail, l'importance de la qualité des images est relevé au vu des meilleures performances de leur modèle sur des bases de données d'images de bonne qualité, comme Messidor-2 [35]. En effet, ils évaluent la proportion des images de qualité de la base de données Kaggle 2015 à 75%. Dans ce travail, il est question de classification binaire ayant pour but de distinguer un patient malade d'un patient sain. Ces deux travaux étudient aussi la difficulté des modèles pour classifier certaines classes. En effet, la classe de RD modérée (R2) décrite dans le tableau 1.1 pose beaucoup de problèmes aux modèle car la différence avec les autres classes est parfois subtile.

Notre travail sur l'évaluation de la qualité des images de fond d'oeil a pour but de lier les avantages des méthodes basées sur la segmentation et ceux des méthodes basées sur l'apprentissage machine. Nous utiliserons l'apprentissage profond, permettant une généralisation sur un grand nombre de données afin d'effectuer la segmentation des régions d'intérêt pour obte-

nir des résultats interprétables qui suivent la définition du SDRGS. Ensuite, nous étudierons différents modèles de dépistage et l'utilisation des techniques d'apprentissage profond pour effectuer le dépistage automatique.

CHAPITRE 3 MÉTHODOLOGIE

Le diagnostic automatique de la RD permettrait d'aider les cliniciens dans le dépistage de la maladie, en introduisant des outils de vision par ordinateur. Cependant, il n'existe pas encore de solution déployée en clinique pour le moment.

Dans ce chapitre, nous détaillons les méthodes retenues pour effectuer ce diagnostic. Tout d'abord, nous abordons le modèle d'évaluation de la qualité des images de fond d'oeil. Ensuite, nous présentons les algorithmes réalisés pour effectuer le dépistage automatique. Enfin, nous étudions l'importance de la qualité dans l'entraînement et le test des modèles de dépistage.

Dans ce chapitre, tous les modèles et algorithmes ont été réalisés en Python. La bibliothèque logicielle utilisée pour les modèles d'apprentissage profond est PyTorch [36], les bibliothèques utilisées pour le traitement d'image sont Pillow [37] et OpenCv [38]. Les bibliothèques utilisées pour le traitement des données sont Numpy [39] et Pandas [40].

3.1 Qualité des images

L'objectif de cette section est d'établir un algorithme d'évaluation de la qualité des images de fond d'oeil. Cet algorithme est inspiré du SGDRS et de la manière avec laquelle les cliniciens évaluent les images de fond d'oeil. Ces images peuvent être réparties en deux catégories, utilisables pour le dépistage ou non selon leur qualité. Une image de bonne qualité contient la macula ainsi que les vaisseaux sanguins autour de la fovea. Le but de l'algorithme est de segmenter ces deux régions d'intérêt et ainsi de donner un score interprétable sur la qualité de l'image. Tout d'abord la segmentation de la macula et des vaisseaux est effectuée de manière indépendante, puis un score est calculé sur une région autour de la macula dans l'image où les vaisseaux sont segmentés. Ce score donne une indication sur la qualité de l'image et un seuil peut être déterminé pour classifier l'image comme utilisable ou non.

Deux réseaux U-Nets [14] différents sont utilisés pour la segmentation de la macula et des vaisseaux dans l'image de fond d'oeil. Ces deux modèles sont entraînés de manière indépendante et possèdent des architectures différentes.

L'algorithme 1 et la figure 3.1 présentent l'entièreté de la méthode d'évaluation de la qualité des images.

Un article publié dans la conférence ICIAR 2020 présentant cette méthode est disponible en annexe A.

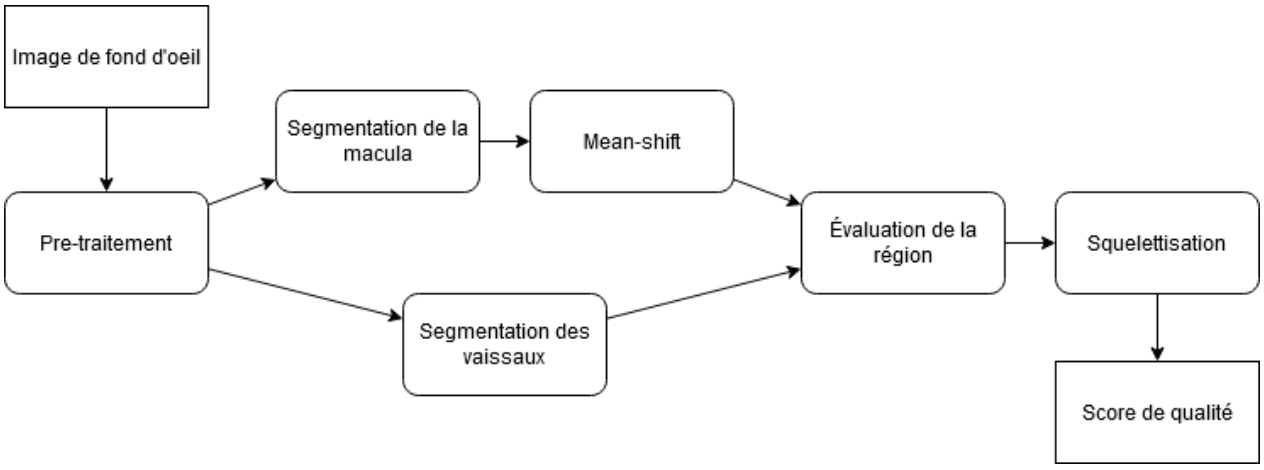


Figure 3.1 Diagramme détaillant le processus complet de la méthode d'évaluation de la qualité des images de fond d'oeil. Les étapes de segmentation de la macula et de segmentation des vaisseaux sont indépendantes. L'étape d'évaluation de la région calcule un score basé sur ces deux segmentations

3.1.1 Pré-traitement

Les deux U-Nets utilisent le même pré-traitement des images de fond d'oeil. Ce pré-traitement consiste à l'application d'une fonction de *Contrast-Limited Adaptive Histogram Equalization (CLAHE)* dans l'espace de couleur LAB. Il s'agit d'une méthode efficace et utilisée couramment pour augmenter la qualité des images de rétine [41]. En effet, la présence de bruit provenant d'une illumination inégale dans les images de fond d'oeil peut heurter les performances des modèles. La fonction de *CLAHE* permet de réduire l'effet de ce bruit en améliorant le contraste de l'image. Ceci permet de mieux percevoir des caractéristiques importantes des images de fond d'oeil comme les vaisseaux sanguins par exemple.

3.1.2 Segmentation de la macula

Un U-Net est utilisé pour segmenter la macula sur l'image après le pré-traitement. L'architecture et la stratégie utilisées sont identiques à celles utilisées par Ronneberger et al. dans [14]. Ce modèle est ensuite entraîné sur 2000 images de la base de donnée Kaggle 2015 manuellement annotée.

Un algorithme de *mean shift* permet ensuite de localiser le centre de la macula et de réduire les erreurs de segmentation. Cet algorithme regroupe les pixels activés par le U-Net selon leur localisation. Les pixels correspondant à la macula sont donc regroupés ensemble et le centre de ce groupe est calculé. Cet algorithme de *mean shift* permet aussi de ne pas tenir

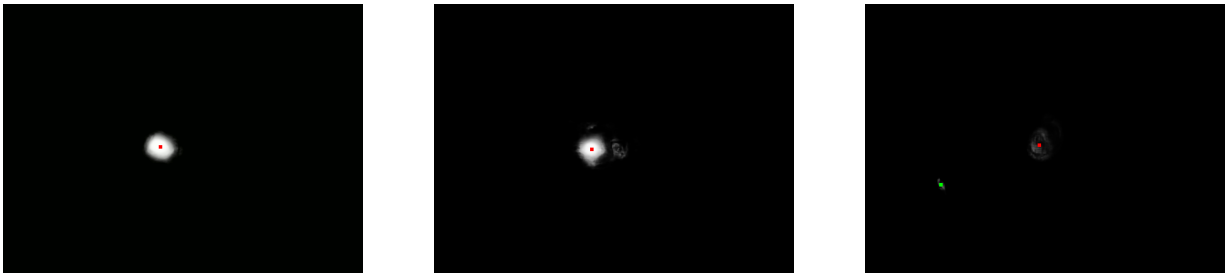
Algorithm 1: Le processus complet de la méthode

```
Result: Score de l'image
image = Image.open(fundusImage)
preprocessed_image = preprocess(image)

unet_macula = load_unet(model_macula)
unet_vessels = load_unet(model_vessels)

macula_segmented = unet_macula(preprocessed_image)
x, y = mean_shift(macula_segmented)

vessels_segmented = unet_vessels(preprocessed_image)

patch = extract_patch(vessels_segmented, x, y)
skeleton = skeletonize(patch)
score = sum(skeleton)

return score
```

compte des erreurs de segmentation du U-Net dans le calcul du centre de la macula car ces erreurs ne seront pas groupées avec la macula. Si la macula ne peut pas être segmentée, l'image est automatiquement classifiée comme inutilisable. La figure 3.2 présente les résultats de l'algorithme sur quelques images de segmentation de macula.

3.1.3 Segmentation des vaisseaux

Le modèle effectuant la segmentation des vaisseaux est aussi un U-Net mais avec une architecture différente de l'originale. Comparé au modèle présenté dans [14], ce modèle possède



(a) (b) (c)

Figure 3.2 Algorithme de *mean shift* permettant de localiser le centre de la macula. En rouge, le centre de la maculé trouvé dans les images. En vert, le centre de groupes de pixels qui ne sont pas la macula

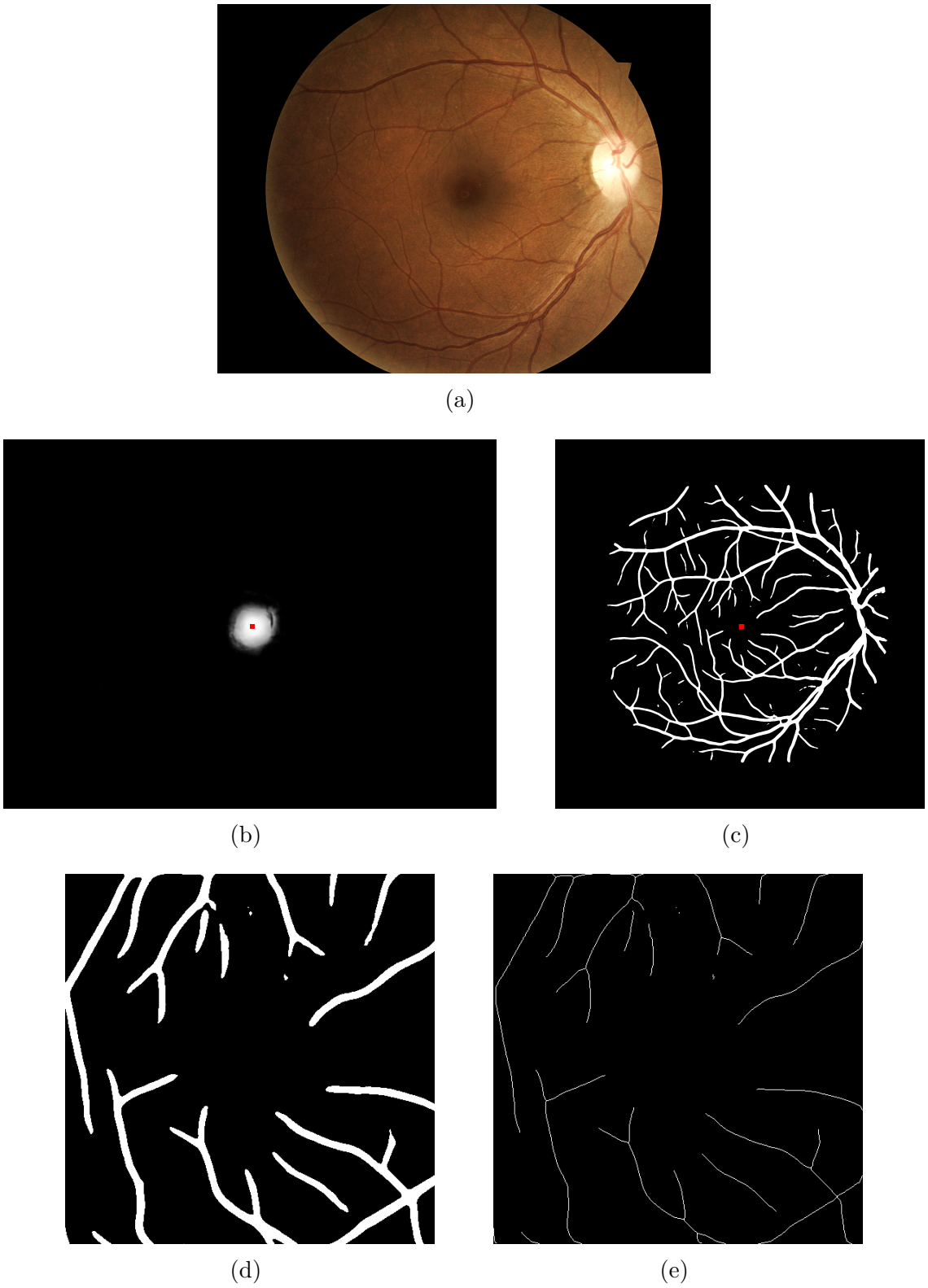
deux fois plus de couches et deux fois moins de caractéristiques par couches. Ceci permet de réduire le sur-apprentissage et élargit le champ récepteur. Ce modèle U-Net est entraîné sur 20 images provenant de la base de données DRIVE [42] et sur 100 images provenant de la base de données Messidor [43]. Ce modèle est entraîné sur 100 *epochs* avec l'optimiseur ADAM [44]. Aussi, les augmentations de données suivantes ont été utilisées : rotations de l'image suivant une loi uniforme entre -180 et +180 degrés, retournement horizontaux de l'image avec probabilité de 0.5, sélection aléatoire d'une région centrée sur un pixel de vaisseau d'une taille de 516x516 pixels, changement de contraste suivant une loi normale d'écart type 0.4 et changement de gamma suivant une loi normale d'écart type 0.15. Ce modèle est utilisé pour segmenter les vaisseaux sur l'image d'origine après pré-traitement. Les résultats sont visibles sur la figure 3.3 (c).

3.1.4 Évaluation de la qualité des images

Les coordonnées du centre de la macula sont utilisées pour extraire une région dans l'image qui contient les vaisseaux. Cette région est centrée sur la macula et contient les petits vaisseaux autour de la fovea. La macula n'est pas visible dans cette image car seulement les vaisseaux sont segmentés.

Il est indiqué dans le SDRGS que cette région doit couvrir une distance d'un diamètre du disque optique depuis le centre de la macula. Afin de décider de la taille de la région, nous avons mesuré manuellement le diamètre moyen du disque optique dans les images de fond d'oeil. Nous avons remarqué que la taille du disque optique varie peu, il est donc possible de fixer le diamètre du disque optique à une valeur constante. Cette valeur correspond à environ 12,5% des dimensions originale de l'image. La taille de la région extraite est donc de 25% de la hauteur et 25% de la largeur de l'image originale, afin de couvrir une distance correspondant au diamètre du disque optique depuis le centre de la macula. Ajouter un moyen de segmenter le disque optique afin d'obtenir une valeur plus précise que cette valeur constante pourrait augmenter la propagation des erreurs car une erreur sur la segmentation du disque optique entraînerait une erreur sur la taille de la région et ainsi entraînerait une erreur sur l'évaluation globale de la qualité de l'image. Les résultats sont visibles sur la figure 3.3 (d).

Ensuite, une squelettisation de la région contenant les petits vaisseaux proche de la fovea est effectuée. Ceci réduit l'impact des vaisseaux larges et met plus d'importance sur le nombre de vaisseaux visibles et leur longueur. Ceci est aussi expliqué par Hunter et al. dans [19]. Cela permet de réduire le nombre de faux positifs puisque les artefacts peuvent parfois être détectés comme des vaisseaux. Avec la squelettisation, leur poids sur le score est réduit significativement. Enfin, le nombre de pixels blancs restants permet d'obtenir le score final



(a)

(b)

(c)

(d)

(e)

Figure 3.3 Détails de l'algorithme de détection de la qualité sur une image de bonne qualité. (a) l'image de fond d'oeil originale. (b) la macula segmentée dans l'image originale. Le centre de la macula est indiqué en rouge. (c) les vaisseaux segmentés dans l'image originale. Le centre de la macula localisé en (b) est indiquée en rouge. (d) la région extraite de l'image (c) centrée sur la macula. (e) le squelette de l'image (d).

de qualité. Les résultats sont visibles sur la figure 3.3 (e).

Ce score est une indication interprétable de la qualité de l'image. Un score nul signifie que la macula n'a pas été segmentée par le premier U-Net ou qu'aucun vaisseau n'est présent dans la région centrée sur la macula. Dans les deux cas, l'image n'est pas de qualité suffisante pour être utilisable dans le dépistage de la RD. Un bon score signifie qu'un grand nombre de vaisseaux longs ont été segmentés autour de la fovea, ce qui implique que la qualité de l'image est suffisante pour se prononcer sur l'avancée de la RD dans l'image. En effet, la qualité d'image autour de la fovea doit être suffisante pour permettre un diagnostic.

Nous pouvons ensuite définir un seuil permettant de classer les images selon leur qualité. La quantité de vaisseaux qui doit apparaître dans la région autour de la fovea n'est pas discuté dans le SDRGS, la valeur est subjective et semble varier selon les images et selon les cliniciens. Afin de définir ce seuil, nous avons fait annoter une base de données de 50 images par un clinicien. Nous avons ensuite déterminé le seuil qui correspondait le plus aux décisions du cliniciens sur cette base de données de test. Le seuil retenu est une valeur de 500 pixels blancs dans l'image du squelette des vaisseaux dans une région autour de la fovea (voir figure 3.3 (e)).

Le seuil permet de réduire le nombre de faux positifs, mais peut ajouter des faux négatifs. Dans notre cas, la précision (équation 2.4) est la mesure la plus importante car plusieurs images d'un même patient sont disponibles et une seule image de bonne qualité suffit pour évaluer la sévérité de la RD. Détecter toutes les images de bonne qualité n'est pas aussi important que d'être certain que les images détectées sont de bonne qualité.

Les résultats du modèle sont discutés dans la section 4.1.

3.2 Modèle de dépistage

Le dépistage automatique de la RD est un problème de classification d'images. En effet, il s'agit d'associer un niveau de maladie à une image. Les niveaux de maladies sont présentés dans le tableau 1.1. Pour simplifier nous parlerons ici des niveaux allant de 0 (absence de RD) à 4 (RD proliférante). Dans cette section, nous allons étudier la conception d'un modèle d'apprentissage profond permettant le dépistage de la RD dans les images de fond d'oeil. Pour cela, nous présenterons les différents modèles conçus pour la compétition Kaggle 2019 et nous expliquerons en détail les approches étudiées.

3.2.1 Kaggle 2019

Nous avons entraîné différents réseaux dans le cadre de la compétition Kaggle APTOS 2019 [31]. Une base de données d'entraînement de 3662 images avec labels était fournie. La figure 3.4 illustre la répartition des classes dans cette base de données. On remarque que la base n'est pas équilibré, la prévalence de la classe 0 est importante.

Lors de cette compétition Kaggle, un système de classement public et classement privé a été mis en place. Une équipe pouvait soumettre jusqu'à 5 soumissions par jour et un score public était dévoilé pour chaque soumission. Ce score public est calculé sur 15% de la base de donnée de test. Les 75% restant sont utilisés pour calculer le score privé qui n'est dévoilé qu'à la toute fin de la compétition, correspondant au score final. Le score public pouvait donc être utilisé comme validation des modèles tout au long de la compétition. Après la fin de la compétition, les scores privés de toutes les soumissions ont été dévoilés. Ces scores seront utilisés afin de comparer les différents modèles utilisés.

3.2.2 Modèle

Pré-traitement

Lors de cette compétition, nous avons étudié l'importance du pré-traitement pour diagnostiquer la RD. Le pré-traitement utilisé est décrit par un membre de la compétition sur le site Kaggle [45]. Il s'agit d'un pré-traitement inspiré de celui du gagnant de la compétition 2015, Benjamin Graham [46]. Ce pré-traitement consiste à soustraire la couleur moyenne locale en utilisant un filtre de flou gaussien. Les images sont aussi coupées pour ne garder que l'oeil et elles sont redimensionnées pour avoir la même taille. La figure 3.5 présente des exemples d'images après pré-traitement.

Augmentation de données

Les transformations que nous avons appliquées sont les suivantes : retournement horizontal et vertical de l'image, translation horizontale et verticale, rotation et agrandissements. La figure 3.6 présente des exemples des différents types d'augmentations utilisés lors de l'entraînement. L'augmentation de données permet de générer de nouvelles données à partir d'une base de données. En général, les performances d'un modèle peuvent augmenter avec le nombre de données différentes, c'est pourquoi différentes base de données peuvent aussi être fusionnés afin de créer une base plus volumineuse. Ces bases de données doivent être similaires pour que la fusion soit bénéfique. La base d'entraînement de la compétition Kaggle Eyepacs 2015 [30]

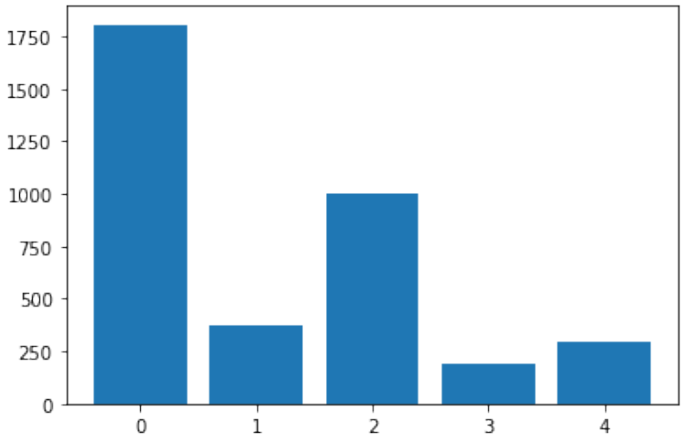


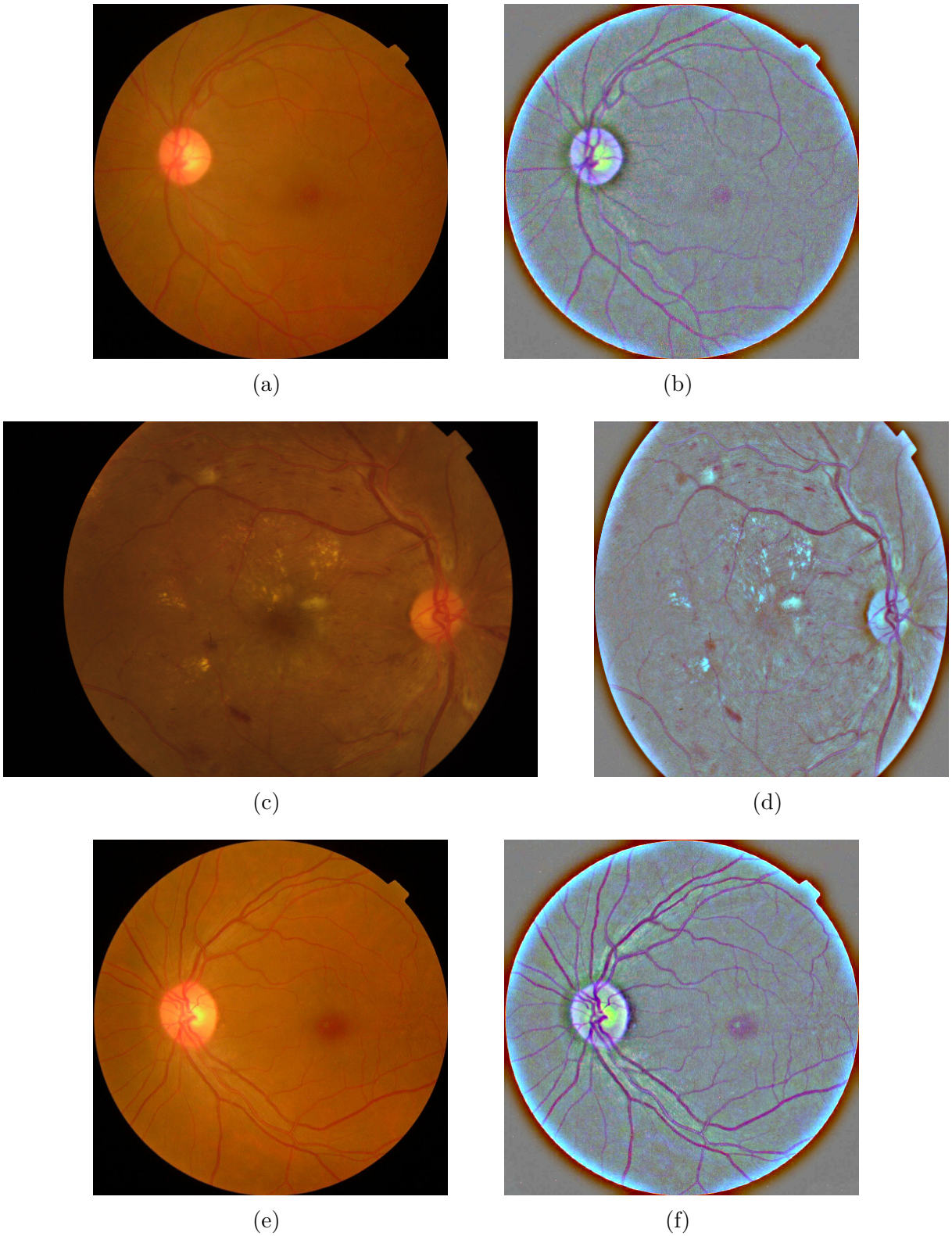
Figure 3.4 Répartition des images selon la sévérité de la RD dans les images de la base de données d'entraînement de kaggle 2019. La proportion des images n'est pas équilibrée, la classe 0 est majoritaire.

compte plus de 30000 images de fond d'oeil similaires à celles de la compétition APTOS 2019. De plus, le système de gradation des niveaux est identique, ce qui fait que la fusion des bases est possible. Augmenter la taille de la base de données d'entraînement ainsi que réaliser des augmentations de données permet au modèle de mieux généraliser et ainsi obtenir de meilleurs résultats.

Régression et Classification

Cette compétition présente un problème de classification, puisqu'il s'agit de répartir les images en différentes catégories discrètes définies par le niveau de RD dans les images. Cependant, une approche de régression peut aussi être abordée. La régression consiste à prédire un nombre réel. Contrairement à la prédiction discrète de la classification, la prédiction de la régression est continue. Ici, la RD est classée en 5 niveaux. Les niveaux sont progressifs (0 à 4) et la régression a pour but de rendre cette progression continue. Les fonctions de pertes associés à ces deux méthodes sont définies dans la section 2.1.4.

Les méthodes de classification et de régression obtiennent des résultats similaires, malgré leur approches différentes. Il est difficile d'interpréter ces résultats, des arguments supportent les deux méthodes. La méthode de classification suit l'approche des cliniciens et du SDRGS en classant les images selon 5 catégories. La méthode de régression tient compte de l'ordre des classes et possède un fonction de perte plus adaptée à la mesure du kappa quadratique (équation 2.1). En effet, l'erreur quadratique moyenne (équation 2.7) et le kappa quadratique mesurent une distance entre la solution et la prédiction. Si une classe 4 est prédite par le

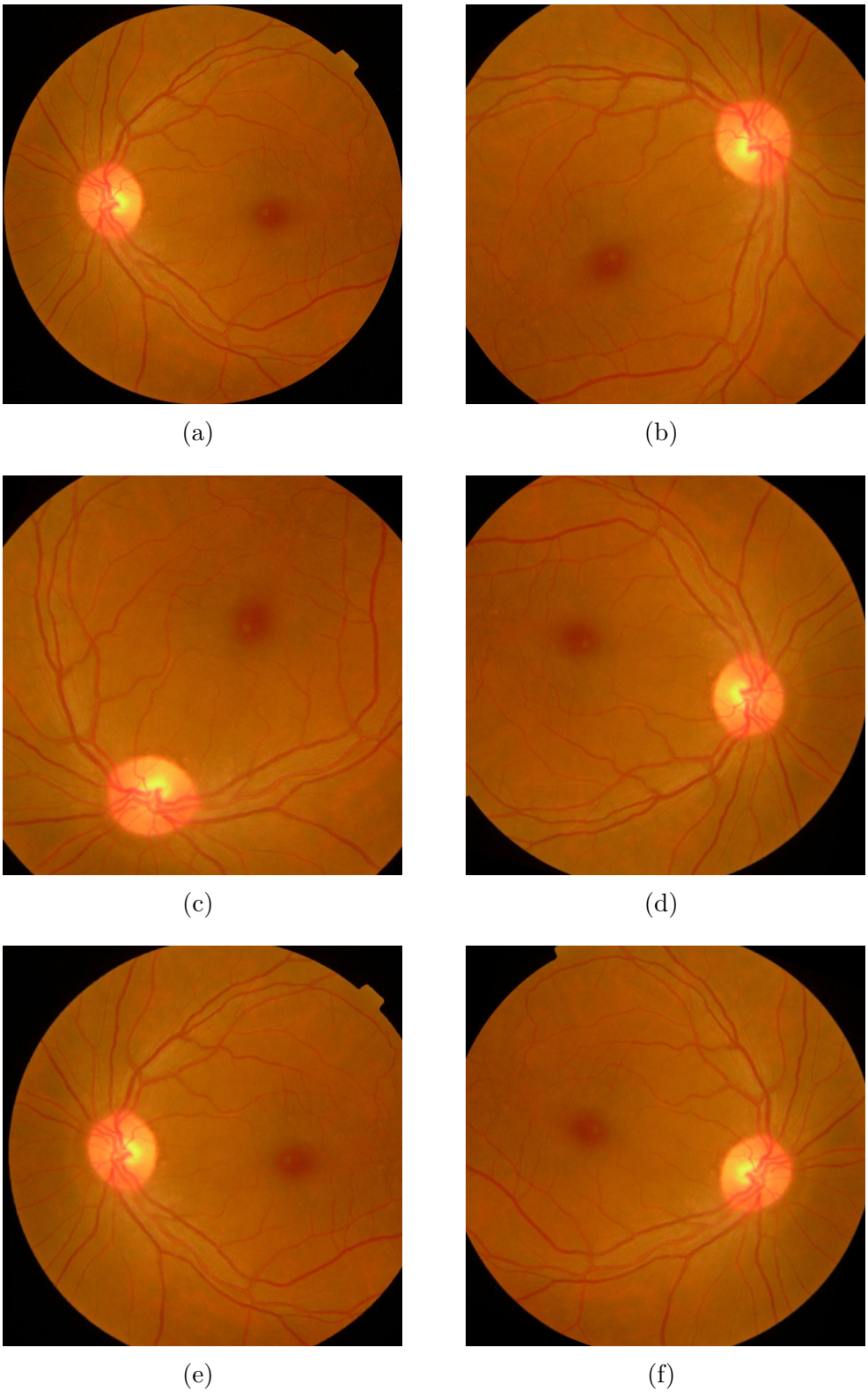


(a) (b)

(c) (d)

(e) (f)

Figure 3.5 Exemple de pré-traitement sur des images de la base de données Kaggle 2019. Pré-traitement inspiré par celui du gagnant de la compétition Kaggle 2015, Benjamin Graham



(a) (b)

(c) (d)

(e) (f)

Figure 3.6 Exemple d'augmentation de données sur une image de la base de données Kaggle 2019. Les augmentations suivantes sont effectuées : retournements horizontal et vertical, agrandissements, rotation, translation

modèle de régression alors que la valeur attendue était 0, l'erreur sera plus importante que si la prédiction était 1. Cette notion de distance n'est pas présente dans la fonction de perte utilisée par la classification (l'entropie croisée, voir équation 2.6). Cependant, la prédiction de la méthode par régression n'a pas de réelle interprétation clinique puisque cette prédiction est un nombre réel qu'il faut ensuite arrondir à une valeur entière comprise entre 0 et 4. Les deux méthodes semblent adaptées au problème, ces approches seront par la suite regroupées dans un modèle ensembliste pour leur complémentarité.

Architecture

Plusieurs types de ResNets [10] et EfficientNets [11] ont été entraînés lors de cette compétition. Ces réseaux convolutifs sont utilisés pour la classification d'images naturelles, sur des bases de données comme ImageNet par exemple. Ceci nous permet d'utiliser l'apprentissage par transfert. Il s'agit d'utiliser les poids des réseaux pré-entraînés sur ImageNet et de les modifier pour effectuer une tâche différente. En effet, les réseaux entraînés sur ImageNet apprennent certaines représentations et caractéristiques des images naturelles. Ces caractéristiques peuvent aussi être utiles pour la classification de la RD, c'est pourquoi entraîner un modèle à partir des poids appris sur ImageNet peut améliorer les résultats.

Les différentes architectures utilisées sont les suivantes : EfficientNet-B3, EfficientNet-B4, EfficientNet-B5, EfficientNet-B6, ResNet-50, ResNet-101.

Apprentissage

Nous avons utilisé l'optimiseur ADAM pour l'apprentissage, ainsi qu'un planificateur de taux d'apprentissage. Ce dernier permet de modifier le taux d'apprentissage lorsque le modèle n'améliore plus son score de validation. Il s'agit donc de réduire le taux d'apprentissage pour permettre au réseau d'effectuer des modifications plus précises et ainsi d'améliorer les résultats.

Aussi, nous avons élaboré une stratégie d'apprentissage qui consiste à d'abord entraîner le réseau sur la base de donnée d'entraînement de Kaggle 2015 composée d'environ 30000 images pour 20 *epochs*. La validation lors de cette première phase d'apprentissage est la base de donnée Kaggle 2019. Le meilleur réseau sur cette première phase d'entraînement sera ensuite utilisé comme départ d'un deuxième entraînement sur la base de donnée d'entraînement de Kaggle 2019. Cette méthode permet au réseau d'apprendre sur une base de donnée importante tout en se spécialisant aux images de la compétition en question.

Méthodes ensemblistes

Nous avons d'abord effectué l'apprentissage des différents réseaux. Ensuite, nous avons appliqué la méthode d'augmentation de données lors de la phase de test à certains réseaux. Ensuite, nous avons regroupé quatre réseaux afin de créer un modèle ensembliste. Ce modèle effectue la prédiction d'une image en calculant l'arrondi de la moyenne des prédictions des réseaux qui le constituent. Afin d'augmenter la variabilité, nous avons regroupé des réseaux avec des architecture différentes (EfficientNet-B3, B4 et B5), des objectifs différents (classification et régression) et une résolution des images d'entrée différente (300, 380 et 456).

Les résultats du modèle sont discutés dans la section 4.2.

3.3 Ajout de la qualité

Nous avons présenté dans la section 3.1 un algorithme d'évaluation de la qualité des images de fond d'oeil et dans la section 3.2.2 un modèle de dépistage. L'objectif de cette section est d'utiliser le score de qualité pour étudier l'impact de la qualité des images de fond d'oeil lors de la phase d'entraînement et de la phase de test, au travers de différentes expériences.

La base de données de Kaggle 2015 est utilisée ici car la base de test est accessible. En effet, la base de test de Kaggle 2019 n'est pas disponible publiquement. La base d'entraînement de Kaggle 2015 contient environ 35000 images de fond d'oeil et la base de test en contient environ 53000. La répartition des données dans la bases de données est présentée en figure 3.7. On remarque encore la proportion très forte des images de classe 0. Nous allons étudier l'impact de la qualité des images dans les modèles de dépistage de la RD en deux parties : lors de la phase d'entraînement et lors de la phase de test.

3.3.1 Phase d'entraînement

Tout d'abord, nous avons filtré les images de mauvaise qualité dans la base de données d'entraînement Kaggle 2015 afin de comparer l'apprentissage de deux modèles identiques sur la base de données originale et la base contenant seulement des images de bonne qualité. Nous avons utilisé des modèles qui ont prouvé leur efficacité lors de la compétition Kaggle 2019 résumé en partie précédente. C'est pourquoi l'architecture choisie pour cette expérience est un modèle EfficientNet-B5. L'objectif du modèle est la classification et il est entraîné sur 25 *epochs* avec un optimisateur ADAM. Nous avons décidé de faire varier certains paramètres de la méthode d'apprentissage afin de mieux comprendre quel facteur permet à la qualité d'impacter les résultats. Ces paramètres variables sont le pré-traitement et les augmentations

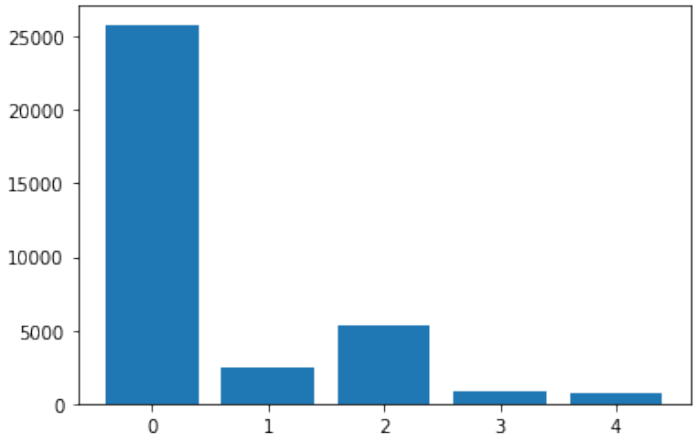


Figure 3.7 Répartition des images selon la sévérité de la RD dans les images de la base de données d'entraînement de kaggle 2015. La proportion des images n'est pas équilibrée, la classe 0 est majoritaire

de données, en plus de la base de données d'entraînement.

Les résultats sont discutés dans la section 4.3.1.

3.3.2 Phase de test

Les performances de différents modèles sont mesurées sur les images de la base de données de test en fonction de leur qualité. Les architectures testées ici sont les mêmes que celles de la section précédente, ainsi que celles présentées dans la méthode ensembliste de la section Kaggle 2019, i.e. EfficientNet-B3, EfficientNet-B4 et EfficientNet-B5. Les réseaux B3, B4 et un B5 ont un objectif de régression tandis qu'un réseau B5 a un objectif de classification. Tous les réseaux sont entraînés pendant 25 *epochs* avec l'optimisateur ADAM sur la base de données d'entraînement Kaggle 2015. Comme indiqué dans la section 3.3.1, certains réseaux sont entraînés seulement sur les images de bonne qualité de la base de données. Certains réseaux sont entraînés avec le pré-traitement de B. Graham [46], et certains avec des transformations d'augmentation de données (retournement et transformations affines).

Les résultats sont discutés dans la section 4.3.2.

CHAPITRE 4 DISCUSSION GÉNÉRALE ET RÉSULTATS

4.1 Résultats du modèle d'évaluation de la qualité

La base de données de validation utilisée pour évaluer les performances du modèle évaluant la qualité des images de fond d'oeil est construite à partir de la base de données utilisée par Fasih et al. dans [21]. Nous avons utilisé 88 images annotées par un clinicien en deux catégories : utilisable et non utilisable. La base de données contient 44 images de chaque classe. La méthode proposée obtient une sensibilité de 100%, une spécificité de 93% une précision de 94% et un rappel de 100% sur cette base. Il existe d'autres bases de données publiques possédant plus d'images, comme EyeQ [47] ou DR2 [48] par exemple, cependant, elles ne suivent pas les critères de qualité définis par le SDRGS. Afin de mesurer au mieux les performances de notre modèle d'évaluation de la qualité des images, nous avons décidé d'utiliser la base de données de Fasih et al. car elle suit les critères de qualité définis par le SDRGS (figure 1.3).

4.1.1 Discussion

L'interprétabilité de notre score nous permet de mieux comprendre les résultats de l'évaluation. Quand une image est classée comme inutilisable, on peut savoir si cette décision provient du manque de vaisseaux autour de la fovea ou si la macula est manquante. Cela nous donne des informations importantes pour mieux évaluer la qualité des images de fond d'oeil. Par exemple, les cliniciens utilisent plusieurs images du même oeil d'un patient pour établir le diagnostic de la DR. Les images possédant un score faible ne sont pas suffisantes pour établir ce diagnostic, mais elles peuvent donner des informations utiles lorsqu'elles sont couplées à d'autres images.

La figure 4.1 présente l'histogramme de la qualité des images, réparties selon leur classe. On remarque une distinction importante entre les deux classes. En effet, le score permet bien de distinguer les images de mauvaise qualité des images de bonne qualité avec le seuil à 500 définit à la section 3.1.4. Cependant, on remarque trois images de mauvaise qualité ayant un score similaire aux images de bonne qualité. Ces images sont faussement détectées comme étant de bonne qualité. La figure 4.2 présente les différentes étapes de l'algorithme pour ces trois faux positifs. On remarque sur les trois exemples que la macula est détectée alors qu'elle n'est pas visible sur les images originales. L'exemple (b) présente des artefacts d'illumination qui entraînent aussi une mauvaise segmentation des vaisseaux. Une absence de segmentation

de la macula aurait permis de correctement classifier ces trois images, soulignant l'importance de la segmentation de la macula.

Nous avons aussi remarqué que 40% des images de mauvaise qualité ont été correctement détectées grâce au modèle de segmentation de la macula. Ceci montre encore l'importance d'une telle étape. Il s'agit d'un critère crucial dans l'évaluation de la qualité des images de fond d'oeil.

4.2 Résultats du modèle de dépistage de Kaggle 2019

Les résultats des différents modèles entraînés pour la compétition Kaggle 2019 sont présentés dans le tableau 4.1. Nous utilisons les résultats obtenus sur la base de test privé de Kaggle 2019 pour comparer les différents modèles. Le kappa de Cohen quadratique décrit par l'équation 2.1 est le score utilisé afin de classer les différents réseaux.

4.2.1 Discussion

Il est notable que les approche régression et classification obtiennent des résultats similaires. De plus, la méthode d'apprentissage semble robuste car elle permet à différentes architectures d'obtenir des résultats similaires. Les réseaux EfficientNet étaient plus facile à entraîner, c'est pourquoi nous les avons largement utilisés dans cette compétition. La méthode ensembliste d'augmentation de données lors de la phase de test permet d'augmenter les performances d'un réseau seul en introduisant de la variabilité. La méthode ensembliste regroupant différents modèles a permis d'obtenir les meilleurs résultats, juste en arrondissant la moyenne des modèles en question. Afin d'augmenter la diversité des modèles utilisés dans cette méthode ensembliste, nous avons fait varier les architectures, les tailles des images et nous avons intégré des modèles de classification et de régression.

On remarque aussi l'importance faible du pré-traitement. Dans notre cas, il ne permet pas d'obtenir de meilleurs résultats. Cela signifie que la qualité des images n'est pas augmentée par le pré-traitement. Le pré-traitement étudié n'est probablement pas assez général et n'est pas adapté à la base de donnée Kaggle 2019. De plus, lorsque les images de cette base sont de bonne qualité, elles sont utilisables sans pré-traitement. Il est possible pour les spécialistes de la rétine de travailler sur des images sans pré-traitement, lorsque la qualité est suffisante. Ils ont à leur disposition l'image avant et après pré-traitement afin de donner leur diagnostic, il est donc probable que les images sans pré-traitement étaient utilisées ici. Aussi, le gagnant de la compétition n'utilise pas de pré-traitement. Ceci souligne l'importance faible des méthodes de pré-traitement étudiées ici.

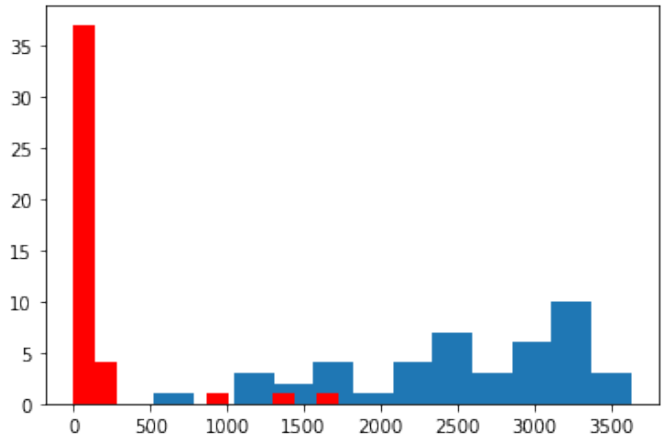


Figure 4.1 Histogramme de la qualité des images en fonction de leur score. En rouge, les images de mauvaise qualité, en bleu les images de bonne qualité. Le score permet de distinguer les deux classes. On remarque cependant trois faux positifs.

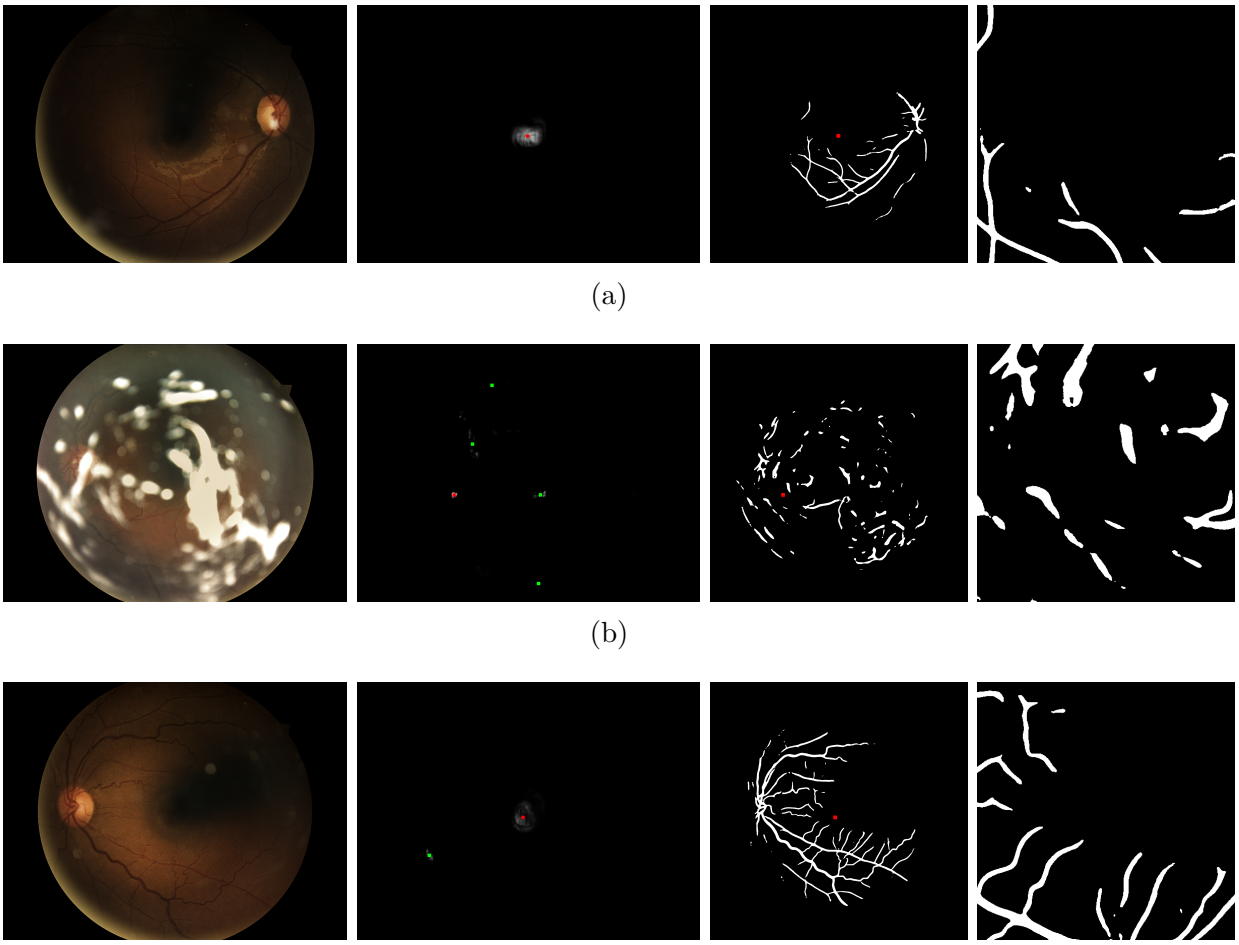
4.3 Ajout de la qualité

4.3.1 Phase d'entraînement

Les performances des modèles sont calculées sur la base de données de test de Kaggle 2015. La qualité est considérée faible lorsque le score de qualité est strictement inférieur à 500. La base de données complète contient 53 576 images, dont 41 934 de qualité haute.

Les résultats obtenus sont visibles dans les tableau 4.2.

On remarque que, lors de l'entraînement, il semble important de conserver les images de mauvaise qualité. En effet, les performances du modèle sont en général meilleures lorsqu'il apprend sur le dataset original. Le meilleur résultat est obtenu lorsqu'aucun pré-traitement n'est appliqué, que des transformations de retournements et affines sont appliquées lors de l'augmentation de données et que l'entraînement est effectué sur le dataset original comportant aussi les images de mauvaise qualité. Ces images de mauvaise qualité semblent quand même contenir des informations importantes pour améliorer les résultats des modèles. Ces informations peuvent être présentes dans les régions éloignées de la fovea par exemple. En effet, si l'image est classée comme étant de mauvaise qualité, cela signifie que la région autour de la fovea n'est pas suffisamment visible. Cependant, les régions plus éloignées peuvent être visible et contenir des informations sur le diagnostic. Ces images peuvent aussi être utilisées par des spécialistes de la rétine lorsque plusieurs image sont disponible pour un même oeil d'un patient. Les images contiennent des informations différentes et elles peuvent être utilisées de manière complémentaire.



(a)

(b)

(c)

Figure 4.2 Faux positifs détectés par le modèle d'évaluation de la qualité des images de fond d'oeil. L'erreur de classification vient principalement du fait que la macula ne devrait pas être détectée

Tableau 4.1 Résultats des modèles obtenus lors de la compétition Kaggle 2019

	Modèle	Input	Preprocess	Augmentations	Training	Détail	Kappa
1	RN-101	224	Aucun	Retournements	2019	Régression	0.787
2	RN-101	224	Graham	Retournements	2015, 2019	Régression	0.885
3	RN-50	224	Graham	Retournements, Affines	2015, 2019	Classification	0.890
4	EN-B5	456	Graham	Retournements, Affines	2015, 2019	Régression	0.894
5	EN-B5	456	Graham	Retournements, Affines	2015, 2019	Régression, TTA	0.910
6	EN-B6	528	Graham	Retournements, Affines	2015, 2019	Régression	0.885
7	EN-B5	456	Graham	Retournements, Affines	2015, 2019	Classification	0.905
8	EN-B5	456	Graham	Retournements, Affines	2015, 2019	Classification, TTA	0.907
9	EN-B4	380	Graham	Retournements, Affines	2019	Régression	0.898
10	EN-B3	300	Graham	Retournements, Affines	2015, 2019	Régression	0.880
11	Ensemble					Model 4, 7, 9, 10	**0.915**

4.3.2 Phase de test

Les performances des modèles sont calculées sur la même base de données de test que la section précédente. Premièrement, le kappa de Cohen quadratique est calculé sur les images dont le score de qualité est faible (kappa 1). Ensuite, nous avons calculé le kappa de Cohen quadratique sur les images dont le score de qualité est élevé (kappa 2). Les résultats sont visibles dans le tableau 4.3.

On remarque que le score sur la base de test contenant seulement des images de bonne qualité est meilleur dans la majorité des cas. Ceci montre que les modèles ont des difficultés à diagnostiquer la RD sur des images de mauvaise qualité. Aussi, l'écart entre les résultats pour les modèles ayant un objectif de régression est bien plus faible que l'écart pour les modèles ayant un objectif de classification. La régression semble être plus robuste au manque

Tableau 4.2 Impact de la qualité des images lors de l'apprentissage

Preprocess	Augmentations	Dataset	Kappa
Aucun	Aucunes	original	**0.766**
Aucun	Aucunes	qualité	0.750
Graham	Aucunes	original	0.695
Graham	Aucunes	qualité	**0.732**
Graham	Retournements + Affines	original	**0.751**
Graham	Retournements + Affines	qualité	0.713
Aucun	Retournements + Affines	original	**0.800**
Aucun	Retournements + Affines	qualité	0.776

de qualité dans les images. La seule différence entre un modèle de régression et un modèle de classification est la fonction de perte. Celle d'un modèle de régression est l'erreur quadratique moyenne, qui tient en compte de l'ordre des classes afin de définir des distances. Ceci permet une marge d'erreur plus grande qu'un modèle de classification. En effet, le diagnostic d'une image de mauvaise qualité est difficile à prédire car certaines caractéristiques de l'image sont peu visibles ou manquantes. L'erreur de prédiction effectuée par un modèle de régression pour la classe 2 est plus importante lorsque la prédiction est 0.3 ou 3.7 que si la prédiction est 1.6 ou 2.4. Ce n'est pas le cas pour un modèle de classification, pour qui l'erreur de prédiction pour la classe 2 lorsque les prédictions sont 0 ou 1 est la même. Un modèle de régression cherche à rapprocher la valeur prédite de la valeur réelle, sans nécessairement l'atteindre pour des images complexes comme des images de mauvaise qualité, alors qu'un modèle de classification cherche à obtenir exactement la valeur réelle, ce qui est difficile pour des images de mauvaise qualité. C'est pourquoi les modèles ayant comme objectif la régression semblent plus robustes au manque de qualité dans les images.

Le modèle ensembliste obtient les meilleurs résultats pour les deux bases de données de test. Ce modèle généralise mieux car il est constitué de différents modèles.

Ce travail sur l'étude de la qualité renforce l'idée que la base de données Kaggle 2015 est complexe puisque la base de test est constituée à plus de 20% d'images de qualité faible. Ceci

Tableau 4.3 Impact de la qualité des images lors du test

	Prétraitement	Augmentations	Dataset	Détail	Kappa 1	Kappa 2
1	Aucun	Aucunes	original	Classification B5	0.748	**0.771**
2	Aucun	Aucunes	qualité	Classification B5	0.714	**0.764**
3	Graham	Aucunes	original	Classification B5	0.664	**0.707**
4	Graham	Aucunes	qualité	Classification B5	0.683	**0.753**
5	Graham	Retournements + Affines	original	Classification B5	0.719	**0.764**
6	Graham	Retournements + Affines	qualité	Classification B5	0.646	**0.744**
7	Aucun	Retournements + Affines	original	Classification B5	0.781	**0.806**
8	Aucun	Retournements + Affines	qualité	Classification B5	0.723	**0.800**
9	Aucun	Retournements + Affines	original	Regression B5	0.801	**0.808**
10	Aucun	Retournements + Affines	original	Regression B4	**0.801**	0.798
11	Aucun	Retournements + Affines	original	Regression B3	0.753	**0.754**
12				Ensemble 7,9,10,11	0.818	**0.822**

a permis de montrer que les modèles sont moins performants sur des images de mauvaise qualité, quel que soit les images d'entraînement. Ceci soulève des remarques intéressantes pour l'entraînement des modèles futurs. En effet, ces modèles peuvent être améliorés en ajoutant cette information de qualité lors de l'entraînement. Sachant qu'une image est de bonne ou de mauvaise qualité, le modèle pourrait s'adapter, en analysant des régions différentes de l'image par exemple. En effet, on a vu précédemment que la région au centre de l'image est très importante dans les images de bonne qualité mais elle est aussi très peu importante lorsque les images sont de mauvaise qualité. Connaître la qualité de l'image en amont pourrait permettre un apprentissage meilleur.

CHAPITRE 5 CONCLUSION

L'objectif de ce travail est d'étudier le dépistage automatique de la RD dans les images de fond d'oeil à l'aide de l'apprentissage profond.

5.1 Synthèse des travaux

Dans la section 3.1, nous avons abordé l'élaboration d'un modèle permettant d'évaluer la qualité des images de fond d'oeil. Ce modèle utilise l'apprentissage profond au travers de deux U-Nets permettant la segmentation de la macula et des vaisseaux sanguins. Ces zones représentent des régions d'intérêt dans l'évaluation de la qualité d'une image. Nous avons ensuite calculé un score de qualité sur une région des vaisseaux autour de la macula. Ce score nous permet d'obtenir 100% de sensibilité et 93% de spécificité sur une base de données de 88 images.

Dans la section 3.2.2, nous avons étudié les performances de différents modèles pour effectuer le dépistage automatique de la RD. Nous avons élaboré un ensemble de modèle obtenant un kappa de Cohen quadratique de 0.915 sur la base de données de test de Kaggle 2019.

Enfin, dans la section 3.3, nous avons étudié l'impact de la qualité des images de fond d'oeil lors des phases d'entraînement et de test des modèles. Dans cette étude, les modèles apprenant sur la totalité des images obtiennent des meilleurs résultats dans la majorité des cas. Aussi, nous avons montré l'importance de la qualité des images lors de la phase de test.

5.2 Limitations de la solution proposée

L'algorithme d'évaluation de la qualité des images de fond d'oeil est basé sur la segmentation de deux régions d'intérêt : la macula et les vaisseaux. La segmentation de la macula est critique, car si elle est défectueuse la sélection de la région dans l'image des vaisseaux sera mauvaise et le score ne sera pas représentatif de la qualité de l'image. Les faux positifs détectés par le modèle présentés sur la figure 4.2 soulignent l'importance du modèle de segmentation de la macula.

L'algorithme de *mean shift* permet de combler certains défauts de la segmentation de la macula. En effet, cet algorithme permet d'améliorer les résultats lorsque plusieurs zones sont extraites par la segmentation de la macula. Cependant, il s'agit d'un algorithme coûteux en temps qui augmente grandement le temps d'exécution de la méthode. Le temps d'exécution

dépend du nombre de pixels segmentés par le modèle de segmentation des vaisseaux. Cet algorithme de *mean shift* peut aller jusqu'à décupler le temps d'exécution, ce qui peut heurter la mise en oeuvre d'une telle méthode.

L'algorithme de dépistage obtient de bons résultats sur la base de données Kaggle 2019 mais ces résultats ne sont pas généralisables sur des bases de données privées d'exemples de cas réels. Une méthode d'apprentissage sur de telles bases de données doit être mise en place afin que le modèle soit utilisable. Ceci peut être expliqué par le fait que les caméras utilisées pour l'acquisition des images sont différentes de celles utilisées par la base de données Kaggle 2019. De plus, plusieurs images d'un même oeil d'un patient sont parfois disponibles dans ces bases de données de cas réels. Le modèle de dépistage pourrait être modifié pour prendre en compte ces multiples images.

L'étude de l'impact de la qualité des images de fond d'oeil sur l'entraînement et le test des modèles de dépistage a été effectuée sur la base de données Kaggle 2015, les résultats peuvent être différents sur d'autres bases de données. Il pourrait être intéressant d'étudier de la même manière l'impact de la qualité des images de différentes bases de données, cela permettrait de mieux comprendre certaines caractéristiques des images de mauvaise qualité et de renforcer les remarques effectuées en section 4.3.

5.3 Améliorations futures

Il est actuellement très compliqué de tester les algorithmes d'évaluation de la qualité des images de fond d'oeil car il existe peu de bases de données de test. De plus, les bases de données utilisent des normes différentes de qualité, ce qui signifie que les performances d'un modèle ne sont pas parfaitement évalués. La création d'une grande base de données permettant d'évaluer les méthodes en fonction des normes du SDRGS pourrait être intéressant. De plus, la base de données utilisée dans ce travail n'est pas beaucoup variée. Les images sont en général de très bonne ou de très mauvaise qualité. Ceci manque d'exemples qui constituent des représentations plus fidèles de la réalité. Les images de qualité moyenne, comme des images floues, des images non centrées sur la macula, ou des images ayant des artefacts permettraient de mieux évaluer les performances d'un modèle, et aussi d'aider un modèle à mieux généraliser.

Une amélioration du modèle effectuant la segmentation de la macula permettrait de se passer de l'algorithme de *mean shift* coûteux en temps. Ce modèle pourrait être amélioré en augmentant la variabilité de la base de données d'entraînement. En effet, l'entraînement du modèle est effectué sur 2000 images provenant de la base de données kaggle 2015. Il pourrait

être intéressant d'ajouter plus d'images différentes.

L'ajout de la segmentation du disque optique peut aussi améliorer les performances et l'interprétabilité du modèle. En effet, le disque optique est une région d'intérêt fortement utilisée par les cliniciens lors de l'évaluation de la qualité. Elle est notamment utilisée dans le SDRGS pour situer la macula et les vaisseaux autour de la macula. Ajouter un modèle de segmentation du disque optique pourrait cependant introduire des erreurs qui auront un lourd impact en fin d'algorithme.

Le modèle de dépistage a été entraîné pour classer les images selon l'avancement de la maladie (0 à 4 selon le tableau 1.1). Il pourrait être intéressant d'étudier un modèle effectuant un dépistage plus simple « référable vs non-référable » . Ceci pourrait réduire les erreurs dues aux classes difficiles à discerner. De plus, il pourrait être d'avantage amélioré en utilisant plus de données correspondant aux patients, comme les autres images d'une visite, les informations correspondant à l'autre oeil, ou même l'âge.

Aussi, le modèle de dépistage prend en entrée des images de taille faible (de 224x224 à 528x528 pixels). Les images originales sont beaucoup plus grandes et permettent de voir des détails autrement invisibles. Ces détails sont très importants pour les cliniciens car il peut s'agir de points hémorragiques, de micro-anévrismes ou d'exsudats durs par exemple. Ces détails sont au coeur du diagnostic de la RD, et ne sont pas visibles à des résolutions plus faibles. Élaborer un modèle prenant un entrée des images de haute résolution permettrait des résultats plus précis, et mieux interprétables.

Enfin, la qualité des images est importante lors de la phase de test. En effet, les modèles ont de meilleurs résultats sur les images de bonne qualité, comme le montre le tableau 4.3. Afin d'améliorer le modèle de dépistage, il pourrait être intéressant d'intégrer le score de qualité lors de l'apprentissage. Ce score pourrait être concaténé aux caractéristiques extraites par le modèle, avant la couche complètement connectées de classification ou de régression. Ce score pourrait aussi être intégré comme pondération de la fonction de perte. La fusion du score de qualité et du modèle de dépistage pourrait ainsi rendre la méthode plus robuste au manque de qualité dans les images de fond d'oeil.

RÉFÉRENCES

[1] Blausen.com staff, "Medical gallery of blausen medical 2014," 2014.

[2] M. Abdullah, M. M. Fraz et S. A. Barman, "Localization and segmentation of optic disc in retinal images using circular Hough transform and grow-cut algorithm," *PeerJ*, vol. 4, p. e2003, 2016.

[3] Early Treatment Diabetic Retinopathy Study standard images, Fundus Photograph Reading Center, University of Wisconsin, Madison, USA.

[4] World Health Organisation, "Global report on diabetes," 2016. [En ligne]. Disponible : https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=21B71A06952630BE3B470260BABD912B?sequence=1

[5] "Scottish diabetic retinopathy grading scheme," https://www.ndrs.scot.nhs.uk/, 2007.

[6] A. Krizhevsky, I. Sutskever et G. E. Hinton, "Imagenet classification with deep convolutional neural networks," dans *Advances in Neural Information Processing Systems 25*, F. Pereira *et al.*, édit. Curran Associates, Inc., 2012, p. 1097–1105. [En ligne]. Disponible : http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[7] J. Deng *et al.*, "Imagenet : A large-scale hierarchical image database," dans *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, p. 248–255.

[8] K. Simonyan et A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv e-prints*, p. arXiv :1409.1556, sept. 2014.

[9] C. Szegedy *et al.*, "Going Deeper with Convolutions," *arXiv e-prints*, p. arXiv :1409.4842, sept. 2014.

[10] K. He *et al.*, "Deep Residual Learning for Image Recognition," *arXiv e-prints*, p. arXiv :1512.03385, déc. 2015.

[11] M. Tan et Q. V. Le, "EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks," *arXiv e-prints*, p. arXiv :1905.11946, mai 2019.

[12] S. Xie *et al.*, "Aggregated residual transformations for deep neural networks," 2016.

[13] Y. Hu *et al.*, "Competitive inner-imaging squeeze and excitation for residual network," 2018.

[14] O. Ronneberger, P. Fischer et T. Brox, "U-Net : Convolutional Networks for Biomedical Image Segmentation," *arXiv e-prints*, p. arXiv :1505.04597, mai 2015.

[15] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, n°. 1, p. 37–46, 1960. [En ligne]. Disponible : https://doi.org/10.1177/001316446002000104

[16] A. Raj, A. K. Tiwari et M. G. Martini, “Fundus image quality assessment : survey, challenges, and future scope,” *IET Image Processing*, vol. 13, n°. 8, p. 1211–1224, 2019.

[17] S. C. Lee et Y. Wang, “Automatic retinal image quality assessment and enhancement,” dans *Medical Imaging 1999 : Image Processing*, K. M. Hanson, édit., vol. 3661, International Society for Optics and Photonics. SPIE, 1999, p. 1581 – 1590. [En ligne]. Disponible : https://doi.org/10.1117/12.348562

[18] M. Lalonde, L. Gagnon et M. Boucher, “Automatic visual quality assessment in optical fundus images,” 01 2001.

[19] A. Hunter *et al.*, “An automated retinal image quality grading algorithm,” dans *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2011, p. 5955–5958.

[20] A. D. Fleming *et al.*, “Automated assessment of diabetic retinal image quality based on clarity and field definition,” *Invest. Ophthalmol. Vis. Sci.*, vol. 47, n°. 3, p. 1120–1125, Mar 2006.

[21] F. Mahnaz, “Retinal image quality assessment using supervised classification,” Masters thesis, École Polytechnique de Montréal, 2014.

[22] L. Abdel-Hamid *et al.*, “Retinal image quality assessment based on image clarity and content,” *Journal of Biomedical Optics*, vol. 21, n°. 9, p. 1 – 17, 2016. [En ligne]. Disponible : https://doi.org/10.1117/1.JBO.21.9.096007

[23] J. M. P. Dias, C. M. Oliveira et L. A. da Silva Cruz, “Retinal image quality assessment using generic image quality indicators,” *Information Fusion*, vol. 19, p. 73 – 90, 2014, special Issue on Information Fusion in Medical Image Computing and Systems. [En ligne]. Disponible : http://www.sciencedirect.com/science/article/pii/S1566253512000656

[24] G. T. Zago *et al.*, “Retinal image quality assessment using deep learning,” *Computers in Biology and Medicine*, vol. 103, p. 64 – 70, 2018. [En ligne]. Disponible : http://www.sciencedirect.com/science/article/pii/S001048251830297X

[25] H. Fu *et al.*, “Evaluation of Retinal Image Quality Assessment Networks in Different Color-spaces,” *arXiv e-prints*, p. arXiv :1907.05345, Jul 2019.

[26] U. R. Acharya *et al.*, “An integrated index for the identification of diabetic retinopathy stages using texture parameters,” *J Med Syst*, vol. 36, n°. 3, p. 2011–2020, Jun 2012.

[27] K. Noronha *et al.*, "Decision support system for diabetic retinopathy using discrete wavelet transform," *Proc Inst Mech Eng H*, vol. 227, n°. 3, p. 251–261, Mar 2013.

[28] R. K. Samala *et al.*, "Mass detection in digital breast tomosynthesis : Deep convolutional neural network with transfer learning from mammography," *Med Phys*, vol. 43, n°. 12, p. 6654, Dec 2016.

[29] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, n°. 7639, p. 115–118, 2017. [En ligne]. Disponible : https://doi.org/10.1038/nature21056

[30] [En ligne]. Disponible : https://www.kaggle.com/c/diabetic-retinopathy-detection

[31] [En ligne]. Disponible : https://www.kaggle.com/c/aptos2019-blindness-detection

[32] X. Guanshuo, "1st place competition summary," https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/108065, 2019.

[33] C. Lam *et al.*, "Automated detection of diabetic retinopathy using deep learning," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2017, p. 147–155, May 2018, 29888061[pmid]. [En ligne]. Disponible : https://pubmed.ncbi.nlm.nih.gov/29888061

[34] A. Rakhlin, "Diabetic retinopathy detection through integration of deep learning classification framework," *bioRxiv*, 2018. [En ligne]. Disponible : https://www.biorxiv.org/content/early/2018/06/19/225508

[35] "Messidor-2 database," http://www.adcis.net/fr/logiciels-tiers/messidor2-fr/.

[36] "Pytorch," https://pytorch.org/.

[37] "Pillow," https://pillow.readthedocs.io/en/stable/.

[38] "opencv," https://opencv.org/.

[39] "Numpy," https://numpy.org/.

[40] "Pandas," https://pandas.pydata.org/.

[41] A. W. Setiawan *et al.*, "Color retinal image enhancement using clahe," dans *International Conference on ICT for Smart Society*, 2013, p. 1–3.

[42] "Drive database," https://drive.grand-challenge.org/.

[43] "Messidor database," http://www.adcis.net/fr/logiciels-tiers/messidor-fr/.

[44] D. P. Kingma et J. Ba, "Adam : A method for stochastic optimization," 2014.

[45] "Kaggle forum preprocessing," https://www.kaggle.com/ratthachat/aptos-eye-preprocessing-in-diabetic-retinopathy, 2019.

[46] B. Graham, "Competition report," https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15801, 2015.

[47] H. Fu *et al.*, "Evaluation of retinal image quality assessment networks in different color-spaces," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, p. 48–56, 2019. [En ligne]. Disponible : http://dx.doi.org/10.1007/978-3-030-32239-7_6

[48] R. Pires *et al.*, "Retinal image quality analysis for automatic diabetic retinopathy detection," dans *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, 2012, p. 229–236.

ANNEXE A ARTICLE 1 : AN INTERPRETABLE DATA-DRIVEN SCORE FOR THE ASSESSMENT OF FUNDUS IMAGE QUALITY

ARTICLE 1 : AN INTERPRETABLE DATA-DRIVEN SCORE FOR THE ASSESSMENT OF FUNDUS IMAGE QUALITY

L'article suivant a été écrit pour la conférence ICIAR 2020. Il résume les travaux effectués sur la qualité des images de fond d'oeil. Cependant, des améliorations ont été effectuées afin de répondre aux problème relevés.

Cet article aborde les deux U-nets présentés dans la section 3.1 ainsi que l'approche d'évaluation du patch. Le U-Net ayant pour but de segmenter la macula a ensuite été amélioré au vues de certains problèmes présentés dans cet article.

An interpretable data-driven score for the assessment of fundus images quality

Youri Peskine[1], Marie-Carole Boucher[2], and Farida Cheriet[1]

[1] Polytechnique Montréal, Montréal QC H3T 1J4, Canada
[2] Hôpital Maisonneuve-Rosemont, Montréal QC H1T 2M4, Canada

Abstract. Fundus images are usually used for the diagnosis of ocular pathologies such as diabetic retinopathy. Image quality need however to be sufficient in order to enable grading of the severity of the condition. In this paper, we propose a new method to evaluate the quality of retinal images by computing a score for each image. Images are classified as gradable or ungradable based on this score. First, we use two different U-Net models to segment the macula and the vessels in the original image. We then extract a patch around the macula in the image containing the vessels. Finally, we compute a quality score based on the presence of small vessels in this patch. The score is interpretable as the method is heavily inspired by the way clinicians assess image quality, according to the Scottish Diabetic Retinopathy Grading Scheme. The performances are evaluated on a validation database labeled by a clinician. This method presented a sensitivity of 95% and a specificity of 100% on this database.

Keywords: diabetic retinopathy · image quality · deep learning · structure-based · data-driven

1 Introduction

Diabetic retinopathy is the leading cause of visual impairment in the working age population as this condition can appear without any symptom. Regular eye examinations are required to enable its detection and treatment. Technicians acquire retinal images of the patient's eyes and retina specialists assess them. Technicians most often take multiple images for each eye of the patient. However, some images may not be used by specialists due to their lack of quality. In the worst case, none of the images meet the quality requirements and specialists cannot grade the images. This can lead to a significant waste of time and resources for technicians, clinicians, and patients. Also, the task of evaluating the quality of a retinal image may have a subjective component as different clinicians or technicians provide different evaluations, based on their experience.

A detailed survey of the image quality assessment methods have been made by Raj et al. [9], dividing methods into three different categories : similarity-based, segmentation-based and machine learning based. Similarity-based methods rely on comparing fundus image features with those a selected set of good

quality images. Segmentation-based methods segment precise structures in fundus images to assess its quality. Blood vessels are the main structure used in segmentation-based methods. For example, Hunter et al. [8] use blood vessels contrast to assess the fundus image quality. Machine learning based methods are data-driven. They learn to classify the data into different categories (e.g. "good quality", "poor quality"). Among these machine learning based methods, we can distinguish two other types of techniques. First, techniques that are based on hand-crafted features and techniques that are based on deep learning features. For example, in [6, 4, 5] the quality of the images are evaluated based on a set of hand-crafted features such as colour, texture or sharpness. These features are not interpretable for clinical purposes, they are typically used to assess the quality of natural images and not fundus images. Deep learning is used in [11, 7] to assess the fundus image quality by extracting features. The features extracted by deep learning based methods are most often not interpretable and described as black boxes.

Our method uses deep learning to segment structures that are relevant to clinicians, resulting to an interpretable score. In this paper, we propose a segmentation-based deep learning method to evaluate the quality of each individual image to help technicians to better assess their quality and retake images when necessary. This new method is based on macula and vessel segmentations, which are regions of interest on retinal images. These regions are extracted using deep learning, but the resulting score is interpretable. This work is inspired by the way clinicians assess image quality in the Scottish Diabetic Retinopathy Grading Scheme [3]. Here, we are focused on evaluating the quality of images used to detect diabetic retinopathy but this work can also be applied to other diseases or condition detection that uses fundus images, such as age-macular degeneration or glaucoma.

2 Method

In this paper, we propose a new method inspired by the quality evaluation in the Scottish Diabetic Retinopathy Grading Scheme. According to this grading scheme, fundus images can be classified as gradable or ungradable based on quality. A gradable image contains the optic disk, the macula and "the third generation vessels radiating around the fovea". Also, the fovea needs to be more than 2 times the diameter of the optic disk from the edges of the image. Here, we will only consider the presence of the macula and the third generation vessels, as they are the most crucial regions of interest to assess the quality of the image. The goal of the method will focus on segmenting these two regions as well as giving an interpretable score to assess the quality of the image. We first segment the macula and the vessels independently in the original image. Then, we compute a score based on a patch around the macula in the vessel segmentation to evaluate the quality of the images. A threshold can be selected to classify the image as gradable or ungradable. Two different U-Net models [10] are used to segment the macula and the vessels. This model has proven to be very efficient

in biomedical image segmentation. The two models are trained independently and have different architectures.

Fig. 1 shows the pipeline of the method on a good quality example. Algorithm 1 presents the entire process.

2.1 Preprocessing

Both U-Net models are using the same preprocessing of the retinal images. This preprocessing consist of applying a Contrast-Limited Adaptive Histogram Equalization (CLAHE) on the LAB color space. This preprocessing is commonly used to enhance images for Diabetic Retinopathy examination.

2.2 Macula segmentation

A U-Net model is first used to segment the macula on the preprocessed image. The architecture and the training strategy are similar to those used by Ronneberger et al. in [10]. The model is then trained on 200 retinal images annotated by retina specialists. This segmentation is used to locate the center of the macula. We compute the mean of the detected pixels to obtain the coordinates of the center of the macula. If the macula cannot be segmented, the image is automatically classified as ungradable. The result corresponds to Fig.1 (b).

2.3 Vessel Segmentation

The vessel segmentation model is also a U-Net model but with a different architecture than the original one. Compared to the model used in [10], this model has twice the number of layers and half the number of features by layers. This prevents overfitting and widens the receptive field. This U-Net model is trained on 20 images from the DRIVE dataset [1] and 100 images from the MESSIDOR dataset [2]. The model is trained for 100 epoch with an ADAM optimizer. We also performed the following data augmentations during the training : rotation, flip and elastic deformations. The model is used to segment the vessels on the original image. The results corresponds to Fig.1 (c).

2.4 Evaluation of the quality of the image

The coordinates of the center of the macula are used to extract a patch from the image containing the vessels. This patch is centered on the macula and contains the small vessels around the fovea. Note that the macula should not be visible in this patch because only the vessels are segmented. The size of the patch should cover approximately 1 diameter of the optic disk from the macula. We decided to set the size of the patch to a constant value of 25% of the original image height and 25% of the original image width because the size of the optic disk is almost constant. Also, by not introducing another model that segments the optic disk, we reduce the propagation of errors, because an error in the segmentation of the

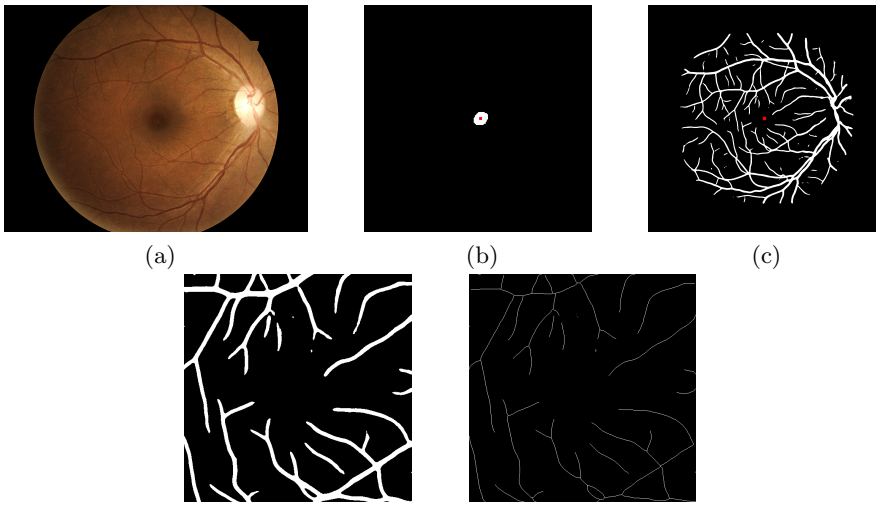


Fig. 1: pipeline of the method on a good quality example. (a) is the original fundus image. (b) is the macula segmented in the original image. The center of the macula is marked in red. (c) is the vessels segmented in the original image. The center of the macula found in (b) is marked in red. (d) is the patch extracted from (c) centered on the macula. (e) is the skeleton of image (d).

Algorithm 1: the entire process of the method

```
Result: Score of the image
image = Image.open(fundusImage);
preprocessed_image = preprocess(image);
unet_macula = load_unet(model_macula);
unet_vessels = load_unet(model_vessels);
macula_segmented = unet_macula(preprocessed_image);
x, y = get_center_of_macula(macula_segmented);
vessels_segmented = unet_vessels(preprocessed_image);
patch = extract_patch(vessels_segmented, x, y);
skeleton = skeletonize(patch);
score = sum(skeleton);
return score;
```

optic disk leads to an error in the global evaluation of the quality of the image. The results corresponds to Fig.1 (d).

Then, we compute a skeletonization on the patch of segmented vessels. This reduces the impact of large vessels and puts more weight on the number of visible vessels and their length. This is also explained by Hunter in [8]. This helps reducing the number of false positive as artifacts can sometimes be detected as vessels. With the skeletonization, their weight on the score is significantly reduced. We then simply count the number of remaining white pixels to obtain the final quality score. The results corresponds to Fig.1 (e).

This score is an interpretable indication of the quality of the image. A null score means that the macula was not found on the image or that the vessels in the patch around the macula were not segmented, resulting in an ungradable image. A good score means that enough lengthy vessels were segmented, resulting in a gradable image. A low score means that only a few vessels were segmented around the macula. Here, we decided to set a threshold to separate gradable and ungradable images based on their score.

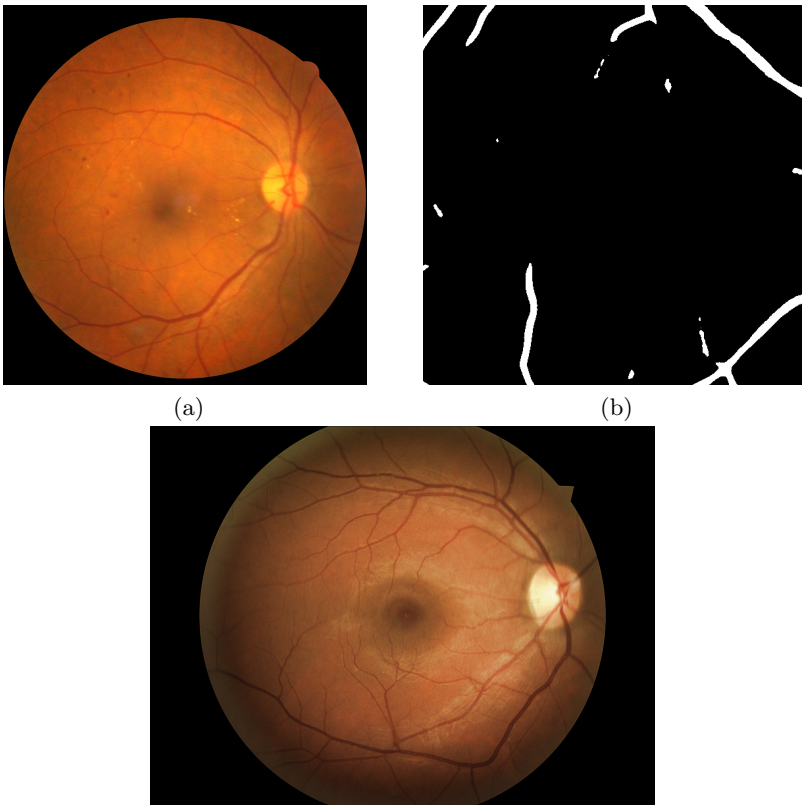
We set the threshold value to match the decisions of a clinician on a training set. The amount of vessels that needs to appear on the image is not addressed in the Scottish Diabetic Retinopaty Grading Scheme., the value is somewhat subjective. The threshold also helps reducing the number of false positives, at the cost of false negatives. In our case, precision is the most important measure because we often have multiple images of a patient's eye, and we only need one gradable image to evaluate the severity of the disease. Detecting all the gradable images is not as much important as making sure that the detected images are indeed gradable.

3 Results and discussion

The validation set used to evaluate the performance of our method is constructed from the dataset used by Fasih in [6]. We used 88 images annotated by a clinician as gradable or ungradable. The dataset contains 44 images of each class. The proposed method obtained a sensitivity of 95% and a specificity of 100% on this validation set.

Only two classification errors were made ; the macula on only two gradable images were not detected. Fig.2 (c) shows one of the images where the macula could not be found by our algorithm, resulting to a misclassification.

The interpretability of our score allows to better understand the output of our method. When an image is classified as ungradable, we can know if this classification is due to the lack of vessels in the patch or if the macula has not been segmented. This gives us important information to better assess the quality of an image. For example, clinicians are using multiple images of the same patient's eye to establish the diagnosis of the diabetic retinopathy. Images with low score are ungradable alone, but they still give some information that can be used by clinicians for the grading of the overall disease, paired with another image.



(a) (b)

(c)

Fig. 2: A low score example, and a false negative example. (a) good quality image where the macula is successfully detected. (b) is the insufficient vessel segmentation around the macula of (a). (c) is a good quality image where the macula has not been detected

Fig.2 (a) and (b) show an example of a good quality retinal image having a low score with our method. This image is classified as gradable by our method, but its score is low and just above our threshold. Fig.2 (b) shows the vessel segmentation around the macula. Compared to Fig.1 (e), very few vessels have been segmented in this example. This means that the image can be gradable, but its quality may not be optimal. If other images of the same patient's eye are available with a better score, they should be prioritized for the diagnostic.

The importance of the macula segmentation is also highlighted in the results of our method. We noted that 80% of the ungradable images were successfully classified due to the macula detection method. This shows how crucial this detection is. The macula is the main criterion to filter out bad quality images.

In this database, all the good quality images were centered on the macula. This is not the case for real life examples. Multiple images are taken for each of the patient's eyes and they are centered on the optic disk as well as on the macula. Our work can be further improved by generalizing this methodology on databases representative of real-life fundus image acquisitions.

4 Conclusion

In this paper, we proposed a new method for evaluating the gradability of fundus images based on the Scottish Diabetic Retinopathy Grading Scheme. This method uses two different U-Net models for the macula and vessel segmentation. The score computed is interpretable and helps understanding the evaluation of quality detection. We achieved a sensitivity of 95% and specificity of 100% while showing the importance of macula segmentation in this methodology. In future works, this method can be used to filter out ungradable images to improve the reliability of deep learning algorithms for diabetic retinopathy grading. Introducing optic disk segmentation is required to further generalize this method by taking into account other relevant information used by clinicians to assess fundus image quality.

5 Acknowledgements

The authors wish to acknowledge the financial support from the CIHR SPOR Network in Diabetes and its Related Complications (DAC) and the department of ophthalmology at the university of Montreal, Quebec, Canada.

References

1. Drive database. https://drive.grand-challenge.org/
2. Messidor database. http://www.adcis.net/fr/logiciels-tiers/messidor-fr/
3. Scottish diabetic retinopathy grading scheme. https://www.ndrs.scot.nhs.uk/ (2007)

4. Abdel-Hamid, L., El-Rafei, A., El-Ramly, S., Michelson, G., Hornegger, J.: Retinal image quality assessment based on image clarity and content. Journal of Biomedical Optics **21**(9), 1 – 17 (2016). https://doi.org/10.1117/1.JBO.21.9.096007, https://doi.org/10.1117/1.JBO.21.9.096007
5. Dias, J.M.P., Oliveira, C.M., da Silva Cruz, L.A.: Retinal image quality assessment using generic image quality indicators. Information Fusion **19**, 73 – 90 (2014). https://doi.org/https://doi.org/10.1016/j.inffus.2012.08.001, http://www.sciencedirect.com/science/article/pii/S1566253512000656, special Issue on Information Fusion in Medical Image Computing and Systems
6. Fasih, M.: Retinal image quality assessment using supervised classification. Masters thesis, École Polytechnique de Montréal (2014)
7. Fu, H., Wang, B., Shen, J., Cui, S., Xu, Y., Liu, J., Shao, L.: Evaluation of Retinal Image Quality Assessment Networks in Different Color-spaces. arXiv e-prints arXiv:1907.05345 (Jul 2019)
8. Hunter, A., Lowell, J.A., Habib, M., Ryder, B., Basu, A., Steel, D.: An automated retinal image quality grading algorithm. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5955–5958 (Aug 2011). https://doi.org/10.1109/IEMBS.2011.6091472
9. Raj, A., Tiwari, A.K., Martini, M.G.: Fundus image quality assessment: survey, challenges, and future scope. IET Image Processing **13**(8), 1211–1224 (2019). https://doi.org/10.1049/iet-ipr.2018.6212
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv e-prints arXiv:1505.04597 (May 2015)
11. Zago, G.T., Andreão, R.V., Dorizzi, B., Salles, E.O.T.: Retinal image quality assessment using deep learning. Computers in Biology and Medicine **103**, 64 – 70 (2018). https://doi.org/https://doi.org/10.1016/j.compbiomed.2018.10.004, http://www.sciencedirect.com/science/article/pii/S001048251830297X