

Titre: Modélisation interprétable du diagnostic de pathologies rétinienues
Title: par apprentissage profond

Auteur: Clément Payout
Author:

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Payout, C. (2023). Modélisation interprétable du diagnostic de pathologies
Citation: rétinienues par apprentissage profond [Thèse de doctorat, Polytechnique
Montréal]. PolyPublie. <https://publications.polymtl.ca/55773/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/55773/>
PolyPublie URL:

**Directeurs de
recherche:** Farida Cheriet
Advisors:

Programme: Génie informatique
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Modélisation interprétable du diagnostic de pathologies rétinienne par
apprentissage profond**

CLÉMENT PLAYOUT

Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Septembre 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

**Modélisation interprétable du diagnostic de pathologies rétiniennes par
apprentissage profond**

présentée par **Clément PLAYOUT**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Guillaume-Alexandre BILODEAU, président

Farida CHERIET, membre et directrice de recherche

Lama SÉOUD, membre

Ismail BEN AYED, membre externe

DÉDICACE

*À mes parents,
pour m'avoir toujours encouragé dans les études,
À mon grand-père,
pour m'avoir donné le goût des sciences...*

REMERCIEMENTS

L'expérience du doctorat est celle de la descente d'un fleuve. On y navigue le long de sinueux méandres, à la recherche de son cap ; il n'est pas rare de s'y perdre, de voguer en rond, repartir en zigzag voire parfois de s'échouer. Pour réalimenter alors le flot clapotant des idées, des affluents impromptus aident à retrouver le sens du courant. Il s'ensuivra des rapides plus ou moins tumultueux interrompus encore par quelques phases de calme plat... Enfin, à l'embouchure et dans un dernier delta rédactionnel, la thèse finit par émerger et se conclut sur ces lignes. Cependant, ce n'est pas du voyage dont je veux parler, mais de la source du fleuve ; en elle réside toute l'énergie du courant et elle ne s'est jamais tarie. Je parle du soutien inconditionnel que j'ai reçu tout au long de ces années. Je n'ai à offrir en échange que ma gratitude la plus profonde. À mes collègues qui, insatisfaits de ce seul statut, sont devenus mes amis. Gabriel, pour la *vraie* clarté de ses analyses, mais surtout le réel plaisir de sa compagnie. Emmanuelle pour la contagiosité de sa bonne humeur, son abnégation à me supporter et pour m'avoir tenu compagnie dans le financement des machines à café. Zacharie pour la rigueur de sa pensée, son enthousiasme pour la recherche, son humour et la démonstration que ces notions sont donc bien compatibles. Philippe, pour ses conseils toujours aussi aiguisés qu'avisés lors des relectures et l'organisation réconfortante de régularité de la vie du labo. À Fantin, Rémy, Antonin, Olivier, Adriana, Marissé, Denise, Catherine, senpai Arnaud... Je ne peux tous vous nommer, mais le roulis aurait été bien moins agréable sans vous. En m'éloignant de l'enceinte du laboratoire : je tiens à remercier mes amis Estelle, Yann, Krystine et Max, sans qui le confinement aurait été bien long, bien morne et plus triste. Léa et Maëlle, pour avoir été des rayons de soleil manquants aux journées parfois ternes de l'hiver. Pierre, pour nos échanges et l'exemple donné de l'importance d'aimer ce que l'on fait. Plus récemment, à mes chers amis Tetiana, Margot et Monsieur Zhorik, pour le plaisir de votre compagnie dans les moments d'isolement ; à quelque chose malheur est bon. Mes remerciements, ils s'adressent bien sûr aussi à Farida, directrice de ce doctorat, pour m'avoir honoré de sa confiance, encore et encore, sans discontinuer ; pour avoir été une oreille attentive quand j'en avais besoin, pour m'avoir apporté un soutien indéfectible, y compris dans mes requêtes les plus incongrues pour aller à l'autre bout du monde. L'expérience du doctorat n'aurait pas eu cette saveur-là sans cette aide, ni sans celle apportée par Santiago et avec nos échanges avec Renaud et Marie Carole.

C'est à ma famille que je veux consacrer ces dernières lignes. À mes frères Simon et Julien, sur qui je sais pouvoir compter malgré la distance qui nous sépare. Et enfin, à mes parents, qui n'ont jamais cessé de croire à la réussite de ce doctorat, à afficher un support sans faille

ni réserve, quelles que soient les circonstances. Ma petite descente de fleuve est bien modeste face à certaines tempêtes de la vie, mais l'exemple de votre résilience a été un phare vers sa destination.

À tous donc, merci d'avoir fait maintenir son cap à cette petite chaloupe.

RÉSUMÉ

Cette thèse traite de la reconnaissance automatique des pathologies rétinienne au sein de deux modalités : la Tomographie par Cohérence Optique (OCT) et le fond d’œil. Les travaux menés s’appuient sur différents types de réseaux de neurones pour la reconnaissance des maladies, mais en s’imposant une condition supplémentaire nécessaire à leur acceptabilité clinique : au delà de la performance d’un algorithme, il faut optimiser sa capacité à produire une prédiction interprétable par un opérateur humain. Plusieurs approches ont été expérimentées. La première consiste à utiliser un modèle de segmentation pour décomposer une image en un ensemble de structures sémantiques significatives pour le diagnostic. Un réseau de neurones convolutif est entraîné à segmenter quatre classes de lésions dans l’imagerie de fond d’œil. Plusieurs bases de données sont utilisées à cette fin, dont la caractérisation détaillée révèle l’hétérogénéité. Or, celle-ci entraîne des contraintes additionnelles sur l’entraînement du modèle dans la mesure où les styles d’annotations divergent fortement. Elle questionne donc sur la capacité de généralisation du modèle. Celle-ci est étudiée sous le prisme de la compatibilité inter-bases et aboutit à la création d’un modèle généraliste capable d’adopter, de manière modulaire, les divers styles de segmentation. L’adaptation est obtenue à partir du concept des attaques adversariales portées sur une sonde d’identification de l’origine d’une image placée au cœur de l’encodeur du réseau. Il en résulte un modèle aux multiples styles de segmentation, configurable par l’utilisateur de façon prédictible, sans ré-entraînement ni même modification de ce dernier.

La segmentation obtenue est convertie en une représentation sous forme de graphe, permettant de préserver la localité des lésions dans un format allégé de l’image. Ce graphe est utilisé pour la gradation automatique de la rétinopathie diabétique, en comparant une architecture existante avec une proposition originale, toutes deux basées sur le principe des GNNs (*Graph Neural Networks*). En limitant la structure du graphe aux lésions segmentées, le lien causal reliant ces dernières au diagnostic est intrinsèquement induit dans le modèle ; ce qui distingue ce dernier de l’état de l’art et favorise la transparence de la prise de décision.

La seconde approche étudiée se focalise sur un modèle récent de classification d’image, le *Transformer*. Un nouveau mécanisme est introduit, basé sur le ré-échantillonnage, d’abord par simple interpolation mais ensuite conditionnel de manière stochastique, de la séquence fournie en entrée. Dans ce second cas, la sélection conditionnelle se fait suivant la carte d’attention autogénérée du modèle. L’algorithme, appelé Attention Concentrée, produit une carte d’attribution locale indiquant les structures lues dans l’image par le réseau de manière précise, contribuant ainsi également à une autre forme d’interprétabilité du diagnostic auto-

matisé.

Pour éprouver les algorithmes et modèles développés, deux applications cliniques seront étudiées : la première, dédiée au suivi de l'évolution longitudinale de la dégénérescence maculaire liée à l'âge au sein d'une cohorte de patients et la seconde au développement d'une plateforme d'aide au diagnostic à des fins de télédepistage.

Ce travail a donné lieu à plusieurs contributions techniques originales. Dans le cadre de la généralisation, l'approche par attaque adversariale permet une adaptation multi-domaines d'une unique architecture conventionnelle, sans modifier ni ré-entraîner celle-ci ; ce qui la distingue des travaux existants dans la littérature. Elle trace une voie vers un entraînement d'un même modèle à partir de données hétérogènes pour lesquelles il n'existe pas nécessairement de domaines source et cible. En termes d'interprétabilité, les travaux sur les architectures Transformers aboutissent à une technique de génération de cartes d'attributions approuvées par les cliniciens. Les Transformers ne permettent pas de manipuler des images de hautes dimensions : l'Attention Concentrée pave le chemin vers de nouvelles architectures se libérant de ces contraintes mémoires.

ABSTRACT

This thesis deals with the automatic recognition of retinal pathologies in two modalities: Optical Coherence Tomography (OCT) and fundus imaging. The work relies on various types of neural networks for disease recognition while imposing an additional condition necessary for their clinical acceptability: beyond the performance of an algorithm, it is essential to optimize its ability to produce predictions interpretable by a human operator. Several approaches have been experimented with.

The first approach consists of using a segmentation model to decompose an image into a set of semantically meaningful structures for diagnosis. A convolutional neural network is trained to segment four classes of lesions in fundus imaging. Several databases are used for this purpose, and their detailed characterization reveals heterogeneity, which poses additional constraints on the model's training due to divergent annotation styles. This raises questions about the model's generalization capability, which is studied through the prism of inter-database compatibility. The result is the creation of a general model capable of modularly adopting various segmentation styles. The adaptation is achieved through the concept of adversarial attacks on an image origin identification probe placed at the heart of the network's encoder. The outcome is a model with multiple segmentation styles, configurable by the user in a predictable manner, without retraining or modification.

The obtained segmentation is converted into a graph representation, preserving the lesions' locality in a lightweight image format. This graph is used for the automatic grading of diabetic retinopathy, comparing an existing architecture with an original proposal, both based on Graph Neural Networks (GNNs). By limiting the graph structure to segmented lesions, the causal link between these lesions and the diagnosis is intrinsically induced in the model, distinguishing it from previous published models and promoting transparency in decision-making.

The second approach focuses on a recent image classification model, the Transformer. A new mechanism is introduced based on resampling, first by simple interpolation and then stochastically conditioned on the self-generated attention map of the model. This algorithm, called Focused Attention, produces a local attribution map indicating the structures read in the image by the network with precision, thus contributing to another form of interpretability in automated diagnosis.

To test the developed algorithms and models, two clinical applications are studied: the first dedicated to monitoring the longitudinal evolution of age-related macular degeneration

within a patient cohort, and the second to the development of a diagnostic support platform for tele-screening purposes.

This work has resulted in several original technical contributions. Regarding generalization, the adversarial attack approach allows for multi-domain adaptation of a single conventional architecture without modifying or retraining it, setting it apart from existing works in the literature. It paves the way for training a single model with heterogeneous data where there may not necessarily be distinct source and target domains. In terms of interpretability, the work on Transformer architectures leads to a technique for generating attribution maps approved by clinicians. Transformers do not handle high-dimensional images, and Focused Attention opens the path to new architectures that are free from these memory constraints.

TABLE DES MATIÈRES

DÉDICACE	iii
REMERCIEMENTS	iv
RÉSUMÉ	vi
ABSTRACT	viii
TABLE DES MATIÈRES	x
LISTE DES TABLEAUX	xiv
LISTE DES FIGURES	xvii
LISTE DES SIGLES ET ABRÉVIATIONS	xxii
LISTE DES ANNEXES	xxiv
CHAPITRE 1 INTRODUCTION	1
1.1 Éléments de la problématique	3
1.2 Objectifs de recherche	5
1.3 Plan de la thèse	5
CHAPITRE 2 CONTEXTE ANATOMIQUE ET CLINIQUE	7
2.1 Structure de l'œil.	7
2.1.1 Anatomie de la rétine	8
2.2 Modalités d'acquisition de l'imagerie rétinienne	9
2.2.1 Le fond d'œil	10
2.2.2 Tomographie par Cohérence Optique (OCT)	12
2.3 Pathologies rétinienne et contexte clinique	15
2.3.1 La rétinopathie diabétique	15
2.3.2 La dégénérescence maculaire liée à l'âge	17
CHAPITRE 3 REVUE DE LITTÉRATURE	20
3.1 Réseaux de neurones dédiés à la vision par ordinateur	21
3.1.1 Réseaux de Neurones Convolutifs	24

3.1.2	Le modèle du Vision Transformer	31
3.1.3	Variantes du modèle	33
3.1.4	Modèles de segmentation sémantique	34
3.1.5	Optimisation et entraînement de modèles	38
3.2	Éléments de la théorie de l'apprentissage statistique	45
3.2.1	Généralisation, sous-apprentissage et sur-apprentissage	45
3.2.2	Distribution multimodale et adaptation de domaines	50
3.3	Comprendre un réseau de neurones	53
3.3.1	Interprétabilité ou explicabilité ?	54
3.3.2	Modèles intrinsèquement interprétables	56
3.3.3	Attributions locales post-hoc	56
3.4	Métriques d'évaluation des performances	63
3.4.1	Scores de classification	64
3.4.2	Segmentation sémantique	69
3.4.3	Un outil dynamique pour appréhender les métriques	72
3.5	Applications aux pathologies rétinienne	74
3.5.1	Segmentation des structures rétinienne dans l'imagerie de fond d'œil	74
3.5.2	Classification de l'imagerie rétinienne	77
3.5.3	Approche statistique de la classification de pathologies	79
3.5.4	Du local au global : approche ascendante de la classification des pathologies par reconnaissance préalable des biomarqueurs anormaux	84
3.6	Discussion	86
CHAPITRE 4 SEGMENTATION DES LÉSIONS RÉTINIENNES DANS L'IMAGERIE DE FOND D'ŒIL		88
4.0.1	Difficultés liées à la segmentation des lésions	89
4.0.2	Variabilité interbases	91
4.1	Problématique	93
4.2	Méthodologie	95
4.2.1	Normalisation des données	96
4.2.2	Conception d'une architecture polyvalente	97
4.3	Résultats expérimentaux	100
4.3.1	Performances comparées entre les architectures de segmentation	100
4.3.2	Étude empirique de la généralisation interbases	101
4.3.3	Adaptation de style par Attaque Adversariale	104
4.4	Gradation automatique de la rétinopathie diabétique assistée par segmentation	118

4.4.1	Expériences préliminaires : le problème de la classification assistée par segmentation	118
4.4.2	Conception et classification d'un graphe rétinien	121
4.4.3	Classification de la rétinopathie diabétique par Réseau de neurones par Graphe (<i>Graph Neural Network</i>) (GNN)	125
4.5	Conclusion	129
CHAPITRE 5 INTERPRÉTABILITÉ D'UN MODÈLE DE CLASSIFICATION PAR ATTENTION CONCENTRÉE		134
5.1	Méthodologie	136
5.1.1	Affinage de l'attention par ajustement du pas	137
5.1.2	L'Attention Concentrée : un ré-échantillonnage itératif conditionnel	137
5.1.3	Entraînement des modèles	139
5.1.4	Bases de données	141
5.1.5	Architectures neuronales	142
5.1.6	Hierarchisation des méthodes d'interprétabilité par attribution	143
5.2	Protocole expérimental	144
5.2.1	Performance de classification	144
5.2.2	Impact du pas adaptatif	146
5.2.3	Évaluation clinique des qualités des cartes d'attribution	147
5.2.4	Performance sur la segmentation des lésions	149
5.3	Discussion	150
CHAPITRE 6 APPLICATIONS CLINIQUES DES ALGORITHMES DÉVELOPPÉS		154
6.1	Recherche de marqueurs prédictifs de la conversion de la dégénérescence maculaire liée à l'âge (DMLA)	154
6.1.1	Mises en application des algorithmes et expérimentations	155
6.2	Considérations sur un hypothétique déploiement clinique	162
CHAPITRE 7 DISCUSSION GÉNÉRALE		167
7.1	Limites des travaux menés	167
7.1.1	Segmentation sémantique, adaptation de domaines et classification par graphes	167
7.1.2	Interprétabilité des modèles et Attention Concentrée	169
CHAPITRE 8 CONCLUSION		171
8.1	Contributions principales	171

8.2 Recommandations	173
RÉFÉRENCES	175
ANNEXES	203

LISTE DES TABLEAUX

Tableau 2.1	Échelle de sévérité de la Rétinopathie Diabétique (RD) telle que définie par l'étude <i>Early Treatment Diabetic Retinopathy Study (ETDRS)</i> .	18
Tableau 2.2	Echelle de sévérité de la DMLA	19
Tableau 3.1	Matrice de confusion dans le cas binaire	65
Tableau 3.2	Matrice de contingence générique entre deux observateurs pour un problème de classification binaire	66
Tableau 4.1	Composition de chacune des bases de données utilisées. Hémorragie intra-vitréenne (HIV), Hémorragie intra-rétinienne (HIR), Hémorragie pré-rétinienne (HPR), Anomalie Microvasculaire Intra-rétinienne <i>Intraretinal Microvascular Abnormality</i> (IrMA), Néo-vascularisation (NV)	93
Tableau 4.2	Architectures utilisées dans notre étude comparative, accompagnées de leur taille (paramètres et FLOPs)	99
Tableau 4.3	Performances comparatives entre les différentes architectures testées. Les modèles sont entraînés et testés sur la base Indian Diabetic Retinopathy Image Dataset (IDRiD) (séparée en deux ensembles suivant les règles de la compétition). En bleu sont indiqués nos scores les plus hauts, en gras, ceux de l'état de l'art.	102
Tableau 4.4	Performance interbases : une même architecture est entraînée sur les différents sous-ensembles des données d'entraînements, dépendamment de leur base d'appartenance.	105
Tableau 4.5	Illustration du phénomène de la sursophistication d'un modèle sur un style d'annotation. Chaque ligne représente un ensemble de test, chaque colonne un modèle. $\mathcal{B}^{(I)}$ représente la base IDRiD, $\mathcal{B}^{(R)}$ la base RETLES. La dernière colonne illustre le paradoxe du modèle combiné, capable d'adapter son style à la base de test.	107
Tableau 4.6	Pour identifier le marqueur de base de données dans l'image, des perturbations géométriques ou colorimétriques sont appliquées à celle-ci. Ces perturbations n'ont pas d'effet sur le style de segmentation choisi par le modèle. Marginalement, le niveau de compression peut avoir un léger effet sur $\mathcal{B}_\star^{(I)}$	110

Tableau 4.7	L'attaque adversariale de la sonde de classification modifie de manière imperceptible à l'œil nu l'image d'entrée. En revanche, elle permet de drastiquement modifier la segmentation pour adopter un style au gré de l'utilisateur.	115
Tableau 4.8	Illustration de l'invariance des performances de classification suivant le modèle de segmentation utilisé. La colonne $\emptyset$ correspond au cas où les caractéristiques de segmentation sont mises à zéro.	121
Tableau 4.9	Performances de gradation de la RD obtenues en testant différentes combinaisons du modèle de segmentation et de classification	122
Tableau 4.10	Performances de classification (gradation en 5 classes) de la rétinopathie diabétique) à partir du graphe de lésions issu des différents modèles de segmentation. Deux architectures de classification sont comparées. On mesure le coefficient κ de Cohen quadratique sur deux ensembles d'évaluation distincts.	127
Tableau 4.11	Performances de classification binaire dans un contexte de détection de la rétinopathie diabétique.	127
Tableau 4.12	Performances obtenues après conversion adversariale vers le style d'annotation de RET-LES et IDRID. Pour cette dernière, nous donnons également les scores avec un graphe élagué des lésions de moins de 64 pixels.	131
Tableau 5.1	Performances comparées entre les modèles de référence de types Réseau de Neurones Convolutif (<i>Convolutional Neural Network</i>) (CNN)s et les modèles Vision Transformer (ViT) pour la classification des OCTs de la base UCSD. Le nombre de paramètres (en millions) et le nombre approximatif d'opérations élémentaires d'addition/multiplications (<i>Multiply-Add Count</i> (MAC)) pour l'inférence sur une image sont donnés à titre indicatif sur la complexité des modèles.	145
Tableau 5.2	Performances comparées sur la classification des OCT de la base Hôpital Maisonneuve-Rosemont (HMR).	146
Tableau 5.3	Performances comparées sur la classification des images de fond d'œil pour deux régimes de données.	147
Tableau 5.4	Évaluation quantitative des performances de segmentation de lésions en utilisant les annotations manuelles disponibles pour les différentes modalités.	151

Tableau C.1	Analyse des performances du modèle par lésion avec et sans prétraitement des données. Il en ressort des résultats très contrastés vis-à-vis de la pertinence de ce prétraitement. En particulier, pour deux modèles et sur deux lésions distinctes (résultats indiqués en rouge), le prétraitement a entraîné un échec complet (et inexplicable) de la segmentation.	212
-------------	---	-----

LISTE DES FIGURES

Figure 1.1	Nombre d'articles de revues recensés sur <i>Google Scholar</i> avec les mots clés « <i>machine learning, retina</i> ».	2
Figure 3.1	Fonctions non-linéaires usuelles placées en sortie de neurones	22
Figure 3.2	Illustration de la notion de champs visuel réceptif obtenu par la mise en séquence de deux couches convolutives de masque 3×3 . Pour obtenir la valeur du pixel rouge, il faut prendre les 3×3 pixels de l'image de la couche précédente, eux-même tous obtenus en utilisant les 5×5 pixels de la couche d'entrée. On peut donc parler de champs visuel réceptif de 5×5 pour la troisième couche.	27
Figure 3.3	Illustration de l' <i>Inception block</i> proposé par Szegedy et al. pour l'architecture GoogLeNet [1]	28
Figure 3.4	Connexion résiduelle proposée par He et al. pour l'architecture ResNet [2]. Les convolutions 1×1 placées en entrée et sortie sont optionnelles : elles permettent la projection des cartes de caractéristiques vers des espaces aux dimensions concordantes si nécessaire.	30
Figure 3.5	Représentation d'une architecture ViT utilisée sur la classification d'une image OCT. Le modèle se compose d'une succession de N couches d'encodage. Le jeton de classification de la dernière couche est fourni en entrée d'une régression linéaire dont la sortie représente la décision du modèle.	33
Figure 3.6	Illustrations des modules utilisées dans l'architecture DeepLabV3 [3] .	36
Figure 3.7	Architecture originale du U-Net	37
Figure 3.8	Définition de la matrice de confusion dans le cas multi-classes. Les images sont issues de captures d'écran prises sur notre outil d'interprétation des métriques de classification. Cet outil adopte une convention différente (transposée de la matrice de confusion). Les notions de Vrai Positif (<i>True Positive (TP)</i>), Vrai Négatif (<i>True Negative (TN)</i>) dépendent de la classe considérée.	68
Figure 3.9	Exemples typiques de courbes ROC (et AUROC associés) pour différents types de modèles	70
Figure 3.10	Évolution de l'indice de Jaccard et du coefficient de Sørensen-Dice sur trois cas de figure théoriques	72

Figure 3.11	Outil interactif de visualisation des métriques de segmentation. L'utilisateur se voit présenter deux canevas sur lesquels il peut dessiner (à gauche, la vérité terrain et à droite, la « prédiction »). Les métriques sont actualisées en fonction des tracés pour illustrer leur évolution. Plusieurs scénarios pré-établis sont configurés pour illustrer les forces et les faiblesses de différentes métriques implémentées.	73
Figure 4.1	Images issues de la base de données RET-LES. L'annotations des lésions est indiquée en surimpression (rouge : hémorragies, vert : exsudats, violet : microanévrismes, orange : nodules cotonneux).	90
Figure 4.2	Images présentées avec leurs annotations manuelles et organisées par colonne suivant leur base d'appartenance. Que ce soit dans les couleurs, le découpage de la région d'intérêt ou les annotations, on constate une grande variabilité inter et intrabase.	92
Figure 4.3	Représentation des distributions jointes de la taille moyenne et du nombre de lésions par images pour chaque classe. Les croix représentent les centroïdes de chaque distribution. Pour pouvoir représenter les distributions sur un même graphique, il a fallu choisir une échelle logarithmique, il y a donc plusieurs ordres de grandeur séparant les centroïdes. En comparant les bases de données, on met ainsi en évidence la disparité des styles d'annotations entre chacune.	94
Figure 4.4	Performance de généralisation par lésion. Suivant l'axe des abscisses sont indiquées les bases d'entraînements, en ordonnées, le score par base d'évaluation. Nous adoptons la métrique Aire sous la Courbe (<i>Area Under the Curve</i>) (AUC) sous la courbe Précision/Sensibilité.	105
Figure 4.5	Résultats qualitatifs de segmentation interbases. Chaque ligne représente un modèle entraîné sur une base. La dernière ligne correspond à la vérité terrain correspondante. Seules les frontières des structures segmentées sont représentées. Cotton-Wool Spot (CWS) , Exsudats (EX) , Hémorragies (HEM) , Micro-anévrismes (MA)	106
Figure 4.6	Modèle de sonde utilisé pour la classification de l'origine des données. Les flèches pointillées indiquent les différents points de connexion, mais il s'agit bien de quatre sondes distinctes, bien que basées sur la même architecture.	111
Figure 4.7	Performance de classification de l'origine des images ($\mathcal{B}^{(i)}$) suivant la couche d'encodage sondée dans le modèle de segmentation.	112

Figure 4.8	Matrice de confusion obtenue à partir des prédictions de la sonde située dans l'avant-dernière couche d'encodage du modèle de segmentation. .	112
Figure 4.9	Pourcentage d'images des bases $\cup_i \mathcal{B}^{(i)}$ classées T par la sonde après conversion $\cup_i \mathcal{B}^{(i)} \Rightarrow \mathcal{B}^{(T)}$. Cette proportion correspond au taux d'images ayant trompé la sonde après conversion vers la cible T visée.	116
Figure 4.10	Cette matrice représente le score mIoU obtenu en comparant les prédictions réalisées par chacun des modèles spécialisés $\mathcal{M}_{(j)}$ et celles réalisées par le modèle $\mathcal{M}_S$ après conversion adversariale des images d'entrées vers chacune des bases.	117
Figure 4.11	Dans le cadre d'une étude préliminaire sur la combinaison des informations de segmentation et de classification. Trois approches ont été expérimentées, basées sur différentes stratégies de combinaisons des deux réseaux	120
Figure 4.12	Les différentes étapes mises en jeu dans l'architecture PointNet++, adaptée à la classification du graphe de lésions. Les étapes de 4.12c à 4.12e représentent les opérations au cœur d'une couche du réseau, répété K fois (selon l'architecture)	124
Figure 4.13	Effet du filtrage des plus petites structures segmentées par $\mathcal{M}[\mathcal{B}^{(I)}]$ sur la classification par graphe.	128
Figure 4.14	Exemples de graphes extraits à partir d'un même modèle $\mathcal{M}_S$ avec ou sans conversion adversariale des images vers IDRID ou RETLES. Sur ces images, chaque nœud est connectée à ses $k = 3$ plus proches voisins, mais elle est simplement indicative. Dans nos expériences de classification, $k = 24$	130
Figure 5.1	Mécanisme illustré de l'Attention Concentrée	139
Figure 5.2	Exemples de la reconstruction de la carte d'attribution haute résolution par Attention Concentrée. Chaque colonne représente un pas différent de l'opérateur de projection convolutif. L'avant dernière et la dernière colonnes représentent respectivement l'attribution finale et la vérité terrain des lésions contenues dans l'image.	140
Figure 5.3	Capture d'écran de l'interface de la plateforme développée dans le cadre du sondage sur l'interprétabilité des réseaux neuronaux mené auprès d'experts médicaux.	145

Figure 5.4	Effet du pas adaptatif sur les performances des modèles pour chaque modalité. Le point carré représente le pas initial du modèle (correspondant au pas lors de l'apprentissage). Le point en forme d'étoile indique le pas correspondant à une performance maximale.	148
Figure 5.5	Scores attribués par les experts sondés concernant la qualité de différentes méthodes d'attribution. Les notes sont données de 1 à 5	149
Figure 5.6	Scores d'interprétabilité sur l'OCT par maladie.	150
Figure 5.7	Analyse des cartes d'attributions issues de différents algorithmes sur des images OCT tirées de la base AROI.	152
Figure 5.8	Analyse des cartes d'attributions issues de différents algorithmes sur des images de fond d'œil tirées de la base IDRID.	153
Figure 6.1	Lien vers le post de blog détaillant les résultats des expériences liées à la DMLA.	156
Figure 6.2	Accord au sens du κ quadratique de Cohen entre les prédictions réalisées par le même réseau sur les images Zeiss ($6 \times 6\text{mm}^2$) et Spectralis.	158
Figure 6.3	Acquisitions faites sur un même patient de la cohorte HMR comparées à un B-scan de la base publique UCSD sur laquelle sont entraînés nos modèles.	159
Figure 6.4	Attention Concentrée calculée sur les patients de la cohorte DMLA. Chaque B-scan correspond à un patient différent et est diagnostiqué comme présentant des drusens par le modèle ViT.	160
Figure 6.5	Segmentation automatique des couches rétinienne dans les images d'OCT structurales. Chaque image de la mosaïque est prise sur un patient différent. <i>ILM</i> : <i>Internal Limiting Membrane</i> , <i>IPL</i> : <i>Inner Plexiform Layer</i> , <i>INL</i> : <i>Inner Nuclear Layer</i> , <i>RPE</i> : <i>Retinal Pigment Epithelium</i> , <i>BM</i> : <i>Bruch Membrane</i> , <i>PED</i> : <i>Pigment Epithelial Detachment</i> , <i>SRF</i> : <i>Subretinal Fluid</i> , <i>IRF</i> : <i>Intraretinal Fluid</i>	162
Figure 6.6	Évolution temporelle de l'arbre vasculaire d'un patient dont l'œil a subi une conversion vers une DMLA humide. Chaque projection vasculaire est réalisée à différentes profondeurs, obtenues grâce à la segmentation menée sur l'OCT structurale correspondante.	163
Figure 6.7	Plateforme de visualisation des prédictions faites par nos modèles. L'utilisateur peut choisir de faire une conversion adversariale du style d'annotations, ce qui influe sur le diagnostic du modèle.	165
Figure 6.8	Illustration de l'importance d'un nœud dans la classification du graphe rétinien et autres statistiques associées consultables par l'utilisateur.	166

Figure C.1	Les différentes étapes du prétraitement appliqué aux images de fond d'œil.	209
Figure C.2	Chaque ligne correspond à une base de données. Le prétraitement proposé vise à homogénéiser les images pour limiter la variabilité inter et intra bases. Il a aussi pour effet d'uniformiser la luminosité et le contraste au sein de l'image.	210
Figure C.3	Histogramme des images de chaque base de données, avant et après prétraitement.	210
Figure C.4	Performance (mIoU) comparée avec et sans pré-traitement sur les différents ensemble de test, en fonction de la base d'entraînement	213
Figure C.5	Règles utilisées pour nos métriques de détection. Les contours pointillés indiquent comment sont interprétées les structures suivant l'intersection de la prédiction et de l'annotation : FP : Faux Positifs (2), FN : Faux Négatifs (1), VP : Vrais Positifs (1)	214
Figure C.6	Métriques de détection. Les abscisses représentent les données utilisées pour l'apprentissage et les différentes barres de couleur le score obtenu sur l'ensemble d'évaluation correspondant.	216
Figure C.7	Performance des différents modèles, entraînés chacun sur une partie de T . Les parties sont triées par ordre croissant de cardinalité, c'est-à-dire par le nombre d'images qu'elles contiennent.	218
Figure C.8	Résultats de différentes méthodes de combinaisons des bases de données. Seul le Dice moyen est fourni ($\frac{1}{5} \sum_{i=1}^5 D(\mathcal{M}, \mathcal{B}_\star^{(i)})$). L'écart-type est représenté en pointillés ($\pm\sigma$).	218

LISTE DES SIGLES ET ABRÉVIATIONS

Pour certaines notions techniques, l'usage de l'acronyme anglais prévaut, y compris dans les milieux francophones. Par souci de clarté, nous utilisons l'acronyme tel que répandu dans l'usage, tout en proposant une traduction française.

Termes médicaux/cliniques	
AREDS	<i>Age-Related Eye Disease Study</i>
AVR	Rapport Artères/Veines (<i>Artery Vein Ratio</i>)
CDR	Ratio Papille/Tête de nerf optique (<i>Cup to Disk Ratio</i>)
CNV	Neovascularisation choroïdale (<i>Chorodial neovascularization</i>)
CWS	Nodules Cotonneux (<i>Cotton-Wool Spot</i>)
DMLA	Dégénérescence Maculaire Liée à l'Age
ETDRS	Early Treatment Diabetic Retinopathy Study
EX	Exsudats
FAF	Fundoscopie par Autofluorescence
FOV	Field of View
HEM	Hémorragies
HMR	Hôpital Maisonneuve-Rosemont
HPR	Hémorragie pré-rétinienne
HIR	Hémorragie intra-rétinienne
HIV	Hémorragie intra-vitréenne
IRM	Imagerie par Résonance Magnétique
IrMA	Anomalie Microvasculaire Intra-rétinienne <i>Intraretinal Microvascular Abnormality</i>
MA	Micro-anévrismes
NV	Néo-vascularisation
OCT	Tomographie à Cohérence Optique (<i>Optical Coherence Tomography</i>)
OMD	Œdème Maculaire Diabétique
RD	Rétinopathie Diabétique
RPE	Epithélium pigmentaire (<i>Retinal Pigment Epithelium</i>)
Termes informatiques	
AUC	Aire sous la Courbe (<i>Area Under the Curve</i>)
AUROC	Aire sous la courbe ROC (<i>Area Under the Receiver Operating Characteristic</i>)

CLAHE	<i>Contrast Limited Adaptive Histogram Equalization</i>
CNN	Réseau de Neurones Convolutif (<i>Convolutional Neural Network</i>)
CRF	Champs Aléatoires Conditionnels (<i>Conditional Random Fields</i>)
CUDA	<i>Compute Unified Device Architecture</i>
GAN	Réseau Génératif Adversarial (<i>Generative Adversarial Network</i>)
GELU	<i>Gaussian Error Linear Unit</i>
GAP	<i>Global Average Pooling</i>
GCN	Réseau de Graphe Convolutif (<i>Graph Convolutional Network</i>)
GNN	Réseau de neurones par Graphe (<i>Graph Neural Network</i>)
GPU	Processeur graphique (<i>Graphical Processing Unit</i>)
HOG	Histogramme des Gradients Orientés (<i>Histogram of Oriented Gradients</i>)
ICP	Itération au Point le plus Proche (<i>Iterative Closest Point</i>)
k-PPV	k-Plus Proches Voisins
LBP	Motifs Binaires Locaux (<i>Local Binary Pattern</i>)
LN	<i>Layer Normalization</i>
LRP	<i>Layer-wise Relevance Propagation</i>
mAP	Sensibilité-Précision Moyenne (<i>mean Average-Precision</i>)
MHA	<i>Multi-head Attention</i>
MSE	Moyenne de l'erreur au carré <i>Mean-Square Error</i>
mIoU	<i>mean Intersection-over-Union</i>
PCA	Analyse en Composantes Principales (<i>Principal Component Analysis</i>)
PMC	Perceptron Multicouches
RELU	<i>Rectified Linear Units</i>
SVM	Machine à Vecteurs Support (<i>Support Vector Machine</i>)
ViT	Vision Transformer

Base de données

ILSVRC	<i>ImageNet Large-Scale Visual Recognition Challenge</i>
DDR	Dataset for Diabetic Retinopathy
FGADR	Fine-Grained Annotated Diabetic Retinopathy
IDRiD	Indian Diabetic Retinopathy Image Dataset

LISTE DES ANNEXES

Annexe A	Modalités d'imagerie de la rétine	203
Annexe B	Bases de données de fond d'œil publiquement accessibles	205
Annexe C	Résultats complémentaires sur la segmentation du fond d'œil	207

CHAPITRE 1 INTRODUCTION

Donner la vue à un algorithme, l'entraîner à déceler ce que l'œil du médecin sait détecter et démocratiser ainsi l'accès aux soins : c'est l'ambition première qui motive la communauté scientifique affairée à la détection automatique des maladies rétinienne. Elle est d'autant plus stimulante que la vision par ordinateur, quand elle a pour sujet d'étude l'œil humain se trouve dans une curieuse mais très utile conjonction. Ce premier objectif axé vers l'assistance au diagnostic se pare parfois d'un second : faire émerger, grâce à la puissance de calcul des algorithmes, des nouveaux modes de compréhension des maladies, de leur pathogenèse ou de leur mécanisme de progression et ainsi accroître le corpus des connaissances scientifiques. Dans un cas comme dans l'autre, cette recherche est rendue passionnante par son positionnement à la confluence de domaines aussi distincts que les mathématiques, l'informatique et la biologie.

Les perspectives d'applications à moyen voire à court terme expliquent à elles seules l'engouement pour ce champ d'étude, en particulier au cours des dernières années. En effet, tous usages confondus, les algorithmes « apprenant » ont récemment fait preuves de progrès spectaculaires. Or, les principales pathologies rétinienne sont en forte augmentation à travers le monde (conséquence notamment du vieillissement de la population). Heureusement les progrès technologiques (comme la numérisation) ont rendu possible la massification de l'imagerie rétinienne. Face à cette double conjonction, l'automatisation de certains aspects du diagnostic est présentée comme une des briques manquantes aux campagnes de dépistage à large échelle. Ce constat est clairement partagé dans le milieu académique : en témoigne l'impressionnante quantité de production scientifique publiée. À ce propos, parler d'effervescence relève de l'euphémisme ; pour s'en convaincre il suffit de regarder la courbe du nombre d'articles liés à la fois à l'apprentissage machine et à la rétine publiés ces dix dernières années (figure 1.1) : la croissance est exponentielle et ne manifeste pour l'instant aucun signe de ralentissement.

Cet engouement est en réalité fort logique, la rétine étant un tissu particulièrement propice à la recherche informatique. En effet, l'imagerie rétinienne est (en général) non invasive et relativement peu onéreuse. Ces propriétés facilitent la collecte et le partage des données à large échelle et donc la communion des efforts de recherche. Plus généralement, un bon nombre de diagnostics de pathologies rétinienne s'appuie sur la lecture et l'interprétation d'image : or la vision par ordinateurs s'est dotée de nombreux nouveaux outils puissants sur la dernière décennie. La prolifération des travaux de recherche consacrés à la détection automatique de pathologies rétinienne n'est donc pas une surprise. Cependant, elle n'est

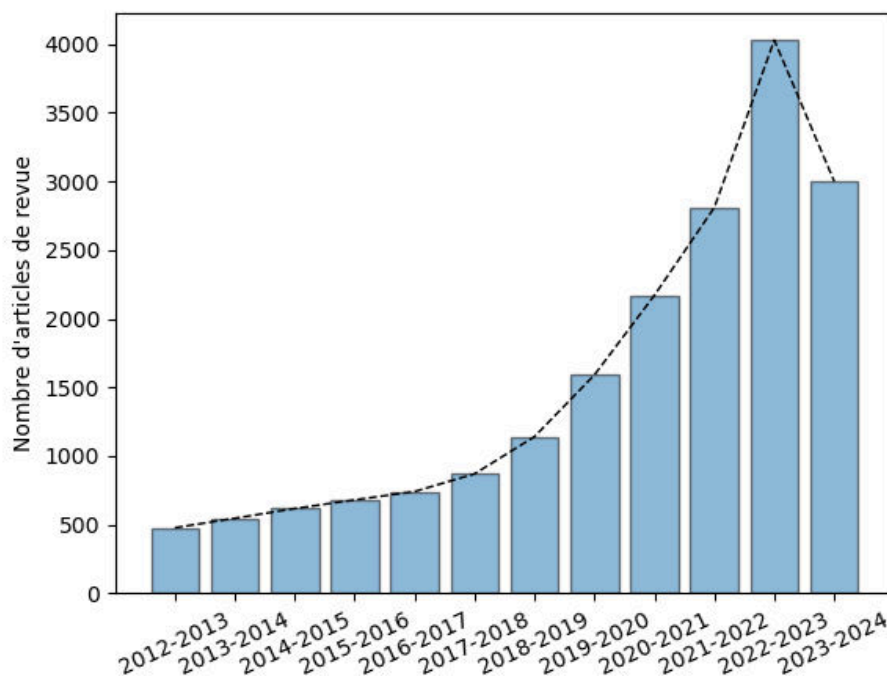


FIGURE 1.1 Nombre d'articles de revues recensés sur *Google Scholar* avec les mots clés « *machine learning, retina* ».

pas -encore- synonyme de déploiement clinique massif. La résolution de tous les obstacles qui jalonnent le chemin entre le laboratoire où est conçu l'algorithme et le patient est loin d'être achevée. Deux problèmes, parmi d'autres, reviennent régulièrement :

- Quelle garantie existe-t-il de l'efficacité d'un algorithme une fois celui-ci sorti de l'univers calibré du laboratoire et appliqué au monde clinique ? Autrement dit, comment garantit-on ses performances ? Cette question s'inscrit dans le contexte de la **généralisation**.
- Comment encourager l'acceptabilité clinique d'un dispositif automatique ? Un algorithme peut-il obtenir la confiance du clinicien ou du patient, de son concepteur ou des autorités de santé ? Ce questionnement est au cœur de tout un pan de la recherche dédié à l'**interprétabilité** des modèles.

C'est sous le prisme de ces deux notions que s'inscrit ce doctorat. Cette recherche a évolué au fil du temps et au gré des collaborations entre les équipes d'ophtalmologie et de recherche de l'hôpital Maisonneuve-Rosemont de Montréal et du laboratoire LIV4D de l'École Polytechnique. Bien que l'application clinique soit toujours en ligne de mire, nos travaux s'inscrivent résolument dans le champ de l'apprentissage machine et le développement de

nouveaux concepts fondamentaux. Pour cette raison, nous n'avons pas cantonné nos expériences à une seule modalité d'imagerie ni à une maladie spécifique. Au contraire, quand cela était possible, nous avons pris le parti d'exploiter les plus grandes collections de données possibles pour éprouver nos propositions théoriques sur les deux thématiques de généralisation et d'interprétabilité. Ce n'est donc pas la conception et le développement d'un seul modèle spécialisé que présente cette thèse, mais les questionnements théoriques et pratiques auxquels un tel modèle se doit de répondre. À ces questionnements et dans la lignée de la littérature, nous proposerons des hypothèses de réponses et leur vérification. Ainsi, traiter de plusieurs maladies et de multiples modalités nous permet d'élargir la portée de nos conclusions. La structuration de ce document suivra cette logique : bien qu'il existe un fil conducteur entre elles, nos contributions méthodologiques peuvent se lire indépendamment les unes des autres. Mais avant de les présenter, le reste de la présente introduction va venir préciser le problème confronté au cours de ce doctorat.

1.1 Éléments de la problématique

L'engouement croissant pour les algorithmes de reconnaissance d'image a trouvé de nombreux échos dans la recherche liée à l'imagerie rétinienne, en particulier pour traiter de la rétinopathie diabétique et de la dégénérescence maculaire liée à l'âge (DMLA). Plusieurs solutions développées par des acteurs privés ou publics ont d'ores et déjà reçu une autorisation de mise sur le marché par les organismes de régulation sanitaire américain et canadien. Ces accréditations sont pour la plupart très récentes (2018 pour la plus ancienne) mais ce marché en forte croissance attire régulièrement de nouveaux acteurs [4, 5]. Tous les logiciels ne sont pas équivalents ; les plus complets permettent d'obtenir une évaluation de la qualité de l'image acquise, la détection de la rétinopathie diabétique (stade modéré ou avancé), la présence d'éléments témoignant de DMLA ou encore des dommages dans la tête du nerf optique indicateurs de glaucome. Si les acteurs qui les développent revendiquent des performances en général très élevées (de pair avec celle d'un médecin), un certain nombre d'interrogations sont régulièrement soulevées quant au fonctionnement des modèles utilisés. La principale critique émise à leur égard concerne leur fonctionnement dit « en boîte noire » : le modèle est capable de prédire correctement un diagnostic, mais l'utilisateur (médecin, patient ou même le concepteur du modèle) est dans l'incapacité de comprendre ce qui a guidé la machine dans son choix. Cet aspect est souvent cité comme un frein à l'adoption clinique de ces systèmes. Plus généralement, la faculté d'un algorithme à expliquer son « raisonnement » est un problème ouvert ; il a donc occupé le cœur de ce doctorat. Or, les maladies rétiniennes que sont la DMLA et la rétinopathie diabétique sont toutes deux diagnostiquées à partir de lésions

locales répondant à des descriptions précises. L'identification de celles-ci (opération appelée segmentation sémantique), préalablement au diagnostic, apparaît donc comme une étape vers une meilleure compréhension du modèle de classification de la maladie. Malheureusement, la segmentation sémantique ne peut se faire qu'avec des données annotées précisément par des médecins et celles-ci sont rares et coûteuses à rassembler. Plus problématiques encore, les protocoles d'annotations varient fortement en fonction des bases de données mises à disposition publiquement. Bien que celles-ci soient très utilisées pour concevoir et valider des modèles, l'effet de cette hétérogénéité est peu analysé. Nos propres travaux sur la segmentation sémantique s'inscrivent donc en partie dans le cadre d'une étude approfondie sur la notion sous-jacente de généralisation.

Le second aspect que nous abordons est celui de la classification des images à partir des cartes de segmentation obtenues. La littérature s'interroge peu sur la relation entre la qualité de la segmentation et la performance de classification ; cette question reste donc à élucider.

Enfin, il existe une alternative interprétable à la segmentation consistant à utiliser directement un modèle de classification tout en explicitant ses mécanismes de fonctionnement internes. Dans une certaine mesure, il est en effet possible d'expliquer le processus de prise de décision d'un modèle. Cette étude relève d'un champ de recherche appelé « interprétabilité des modèles ». S'il existe de nombreuses façons de l'aborder, celle qui nous intéressera a trait à la notion de génération de cartes d'attributions. Ces dernières jouent un rôle similaire à la segmentation : elles délimitent localement les zones d'intérêts dans l'image sur lesquelles se base le modèle pour sa prédiction. La génération de ces cartes a l'avantage de ne pas nécessiter d'annotations locales, précisément celles qui sont coûteuses à obtenir. Malheureusement, les divers algorithmes d'attributions existants sont loin d'égaliser la précision et la résolution obtenues avec une réelle segmentation, elles ne peuvent donc prétendre les remplacer en l'état. Au-delà des performances mesurées sur des métriques (qui peuvent être elles-mêmes problématiques), l'applicabilité des algorithmes développés pose aussi question. Dans un environnement où des solutions logicielles sont déjà commercialisées, quelle place reste-t-il à des outils de recherche qui rivaliseront difficilement sur ces métriques ? Cette question naît du constat que dans la majorité des travaux publiés dans le domaine, les conclusions ne s'étendent guère au-delà des scores de performances obtenus sur des bases de référence. L'utilisabilité de ces algorithmes, que ce soit à des fins de recherche ou de déploiement clinique, est souvent laissée en suspens.

1.2 Objectifs de recherche

Au vu des différents éléments de la problématique, nous avons fixé deux objectifs de recherche. Le premier, dédié au fond d’œil et à la segmentation, s’articule autour des objectifs spécifiques suivants :

- Caractériser les bases de données publiques dédiées à la segmentation de l’imagerie de fond d’œil (plus spécifiquement des lésions).
- Concevoir un modèle de segmentation sémantique automatique capable de détecter les lésions rétinienne sur différentes bases de données tout en atteignant les performances de l’état de l’art.
- Analyser les conséquences de l’hétérogénéité des données sur l’entraînement et la validation du modèle et expérimenter des stratégies permettant de les modérer voire de les contrôler.
- Construire un modèle causal de gradation de la rétinopathie diabétique à partir des cartes de segmentation des lésions.

Le second objectif de recherche explore l’autre aspect discuté dans la problématique, l’interprétabilité des modèles de classification. Nos travaux sont bâtis autour des objectifs suivants :

- Comparer les architectures convolutives traditionnelles aux récents modèles de type autoatmosphériques (plus particulièrement les *Vision Transformer* [6]) à des fins de classification du fond d’œil et de l’OCT.
- Étudier les algorithmes de génération de cartes d’attributions existants et proposer une approche permettant d’affiner celles générées par le modèle ViT.
- Comparer qualitativement et quantitativement les cartes d’attributions. Pour l’aspect quantitatif, nous utiliserons les annotations de segmentation faites par des cliniciens.
- Impliquer les médecins dans l’évaluation de l’utilité des différentes cartes d’attributions et donc proposer un protocole de collaboration permettant de recueillir leur contribution de manière efficace et non-biaisée.

1.3 Plan de la thèse

Le prochain chapitre s’adresse au lecteur peu familier du contexte ophtalmologique qui entoure cette thèse. Nous y synthétisons les principaux éléments qui seront manipulés par la suite, que ça soit les notions élémentaires de l’anatomie rétinienne, des pathologies afférentes et les différentes modalités physiques utilisées pour faire l’acquisition des images que nous manipulerons. Pour contextualiser nos travaux au sein de la littérature existante, le chapitre qui suivra présente l’état de l’art propre à notre domaine. Il s’organise en deux parties : la

première, à vocation généraliste, introduit les notions fondamentales de l'apprentissage par réseaux de neurones. Son contenu s'adresse au lecteur peu au fait des avancées du domaine obtenues lors des dix dernières années. À partir de la section 3.2.2, nous focaliserons notre revue sur les travaux plus récents et plus directement liés à notre domaine. On y couvre, dans l'ordre, les questions de généralisation, d'interprétabilité et de mesure des performances des algorithmes. Enfin, la revue de littérature s'achève sur une présentation des travaux consacrés à l'ophtalmologie utilisant ces différents concepts d'apprentissage.

Les chapitres méthodologiques suivent peu ou prou l'ordre des objectifs de recherche que nous nous sommes fixés. Le chapitre 4 traite de la segmentation sémantique dans l'imagerie de fond d'œil. Nous y proposons une comparaison pléthorique de modèles entraînés sur diverses bases de données, elles-mêmes préalablement caractérisées en termes de qualité d'images et d'annotations. L'hétérogénéité des bases de données y sera étudiée et en particulier, nous y discuterons du comportement inattendu qu'elles peuvent engendrer. Cette observation aboutira au développement d'une stratégie originale de conversion des données dite « adversariale » permettant de changer de style d'annotations sans altérer un réseau entraîné. Dans un second temps, ce même chapitre introduira une nouvelle manière de classer les images de fond d'œil s'appuyant sur une représentation par graphe de celle-ci. Cette contribution originale a pour objectif de quantifier la qualité de la segmentation au travers des performances de classification et nous montrons qu'en cela, elle se distingue significativement des travaux existants.

Le chapitre 5 se consacre à la question de l'interprétabilité ; nous y introduirons pour cela divers modèles entraînés sur des bases de fond d'œil et d'OCT. Un intérêt tout particulier est consacré aux modèles de type ViT. Au cours de nos expériences, la découverte d'un mécanisme de ré-échantillonnage des séquences d'entrées du dit modèle nous amènera à exploiter des capacités insoupçonnées de ce dernier et à améliorer ses performances sans avoir à le réentraîner. Toujours dans l'optique de fournir des diagnostics transparents pour le médecin et en s'appuyant sur cette découverte, nous introduirons le principe d'Attention Concentrée, qui nous permet de générer des cartes d'attributions de hautes résolutions.

Le chapitre 6 propose de revenir sur les algorithmes conçus au cours de la thèse sous forme d'une discussion sur leur applicabilité. Pour cela, nous avons choisi de présenter succinctement diverses expériences cliniques menées au cours de cette thèse utilisant les algorithmes développés. Les méthodologies mises en place et les résultats obtenus n'y sont que brièvement mentionnés, car nous prenons le parti pris de privilégier une discussion sur les forces et faiblesses de nos différentes contributions à la lumière d'applications cliniques bien réelles.

La discussion autour de l'ensemble des résultats obtenus ainsi que des limites de nos contributions concluent ce document, au sein des chapitres 7 et 8.

CHAPITRE 2 CONTEXTE ANATOMIQUE ET CLINIQUE

Pour saisir les différentes composantes de notre problématique, il est nécessaire de revenir sur le contexte dans lequel elle s’inscrit. La recherche sur l’imagerie médicale impose un certain nombre de contraintes spécifiques :

- L’interprétation d’une image ne peut se faire sans un minimum d’expertise sur la structure anatomique imagée. La collaboration avec les médecins est donc souvent indispensable, en particulier pour l’annotation des données qui dépend entièrement de leur contribution.
- Plus pragmatiquement, la collecte des données est en général plus ardue que sur d’autres domaines. Plusieurs facteurs l’expliquent : le respect du consentement éclairé du patient est impératif et l’approbation par un comité d’éthique avant la collecte rajoute des étapes protocolaires. Chaque acquisition est donc coûteuse en temps ou en moyens (pour le patient et les équipes médicales en charge).

À la différence peut-être d’autres applications du *Big Data*, ces contraintes rendent chaque donnée individuelle très précieuse. Il est donc important pour quiconque la manipule y compris le chercheur en informatique de comprendre a minima les bases anatomiques de la structure imagée, les contraintes physiques liées à la modalité d’acquisition et la manifestation des pathologies qu’on souhaite pouvoir détecter. Ce chapitre se consacre donc à détailler ces différents points.

2.1 Structure de l’œil.

L’œil transforme la lumière reçue en influx nerveux retransmis au cerveau. Tout comme un appareil photographique ne peut être réduit à un simple capteur photonique, l’œil est constitué d’un ensemble sophistiqué de tissus venant guider la lumière vers la rétine. Dans son apparence externe et à l’œil nu, le globe oculaire se compose de la sclère (« blanc de l’œil ») et de la cornée. La sclère est une structure résistante, visant à protéger l’œil des dégâts mécaniques. Elle est recouverte (hormis sur sa partie antérieure) par la conjonctive, responsable de la production du mucus intégrant la composition du liquide lacrymal. Sur sa partie antérieure, la sclère est percée pour laisser place à la cornée. Il s’agit d’une membrane circulaire bombée et transparente constituée à 78% d’eau. Les larmes produites par les glandes lacrymales viennent l’alimenter en permanence et sont réparties sur toute sa surface par le battement des paupières. Elle est la principale responsable de la focalisation des rayons lumineux pénétrant l’œil : à elle seule, elle assure près de 80% de la réfraction. Derrière la

cornée se trouve une ouverture, nommée pupille, fonctionnant comme un diaphragme. La largeur de la pupille, contrôlée par les muscles lisses de l'iris, permet d'adapter la quantité de lumière pénétrant l'œil.

Pour assurer une mise au point aussi bien sur des objets lointains que proches, la focalisation doit être dynamique. Cette dynamique est rendue possible grâce au cristallin, petit disque fibreux situé à l'arrière de la cornée. La tension du cristallin est contrôlée par le corps ciliaire, constitué d'un réseau de muscles (muscles ciliaires) et également responsable de la sécrétion de l'humeur aqueuse. L'ensemble cornée/cristallin joue le rôle de doublet optique dont la longueur focale est régulée par la tension appliquée au cristallin. Derrière ce dernier se trouve le corps vitré, où circulent les rayons lumineux avant de converger vers la surface tapissant l'intérieur du globe oculaire. Cette dernière, appelée rétine, est responsable de la conversion électromagnétique du signal lumineux en impulsions nerveuses.

2.1.1 Anatomie de la rétine

La vision humaine est le résultat de l'interaction complexe entre la lumière et des photorécepteurs qui convertissent et envoient les signaux nerveux au cerveau. Le siège de cette interaction se trouve dans la rétine, tissu qui tapisse la surface interne de l'œil. Au cours de l'embryogenèse, la rétine se développe depuis la cupule optique (tête du nerf optique), une excroissance issue du cerveau frontal [7]. De cette tête de nerf optique émergera le réseau vasculaire chargé d'irriguer l'ensemble des couches de la rétine ; il est compris d'artères, de veines et de microvaisseaux nommés capillaires rétiens. En ce qui concerne les couches rétiennes, elles se distinguent par le type de neurones qui les composent. Il en existe cinq classes majeures dans la rétine : les photorécepteurs, les cellules bipolaires, les cellules horizontales, les cellules amacrines et les cellules ganglionnaires. Sur des coupes histologiques, on peut en dénombrer un nombre plus important encore et observer précisément l'organisation en couches de ceux-ci. Radialement (en suivant le chemin de la lumière arrivant dans l'œil), ces couches rétiennes superposées sont :

1. La membrane limitante interne, composée des cellules de Müller.
2. La couche des fibres nerveuses, c'est-à-dire les axones des cellules ganglionnaires.
3. La couche des cellules ganglionnaires, qui contient leurs noyaux.
4. La couche plexiforme interne formée des cellules amacrines
5. La couche nucléaire interne formée des cellules bipolaires
6. La couche plexiforme externe formée des cellules horizontales
7. La couche nucléaire externe, où se situent les photorécepteurs (cônes et bâtonnets).

8. L'épithélium pigmentaire (ou RPE pour *Retinal Pigment Epithelium*), par où se font les échanges nutritifs et apports sanguins entre la rétine et les structures plus profondes.

Sous cette dernière couche se trouve un second tissu, appelé la choroïde. C'est par celle-ci, très vascularisée, que se fait la nutrition de l'iris et des photorécepteurs. Parmi ceux-ci, la distinction entre cônes et bâtonnets permet d'expliquer la vision en couleur et la vision périphérique. Tout comme la plupart des vertébrés, l'être humain possède ces deux types de cellules, les bâtonnets étant près de vingt fois plus nombreux que les cônes (on compte entre 110 et 130 millions de bâtonnets pour environ 7 millions de cônes). Les cônes sont responsables de la perception de la couleur. Ils contiennent différents pigments réagissant respectivement à trois plages de longueurs d'onde en envoyant un influx nerveux au cerveau. Ces longueurs d'onde correspondent à peu près au bleu, au vert et au jaune sur le spectre de la lumière. Les bâtonnets pour leur part ne disposent que d'un type de pigment, sensible aux longueurs d'onde proche du bleu ; ils sont par conséquent incapables de distinguer les couleurs. En revanche, étant plus sensibles, ils restent actifs même en faible luminosité. Chez l'humain, près de 50% des cônes sont situés au sein d'une structure centrale appelée macula, au centre de laquelle se trouve la fovéa, point central concentrant la majorité des cônes.

La macula est donc le point central de la vision. Il s'agit d'une région elliptique de 1.5 mm de large pour 1 mm de hauteur. Au niveau de la fovéola (centre de la fovéa), elle est très mince (130 μm). La région dépressionnaire de la fovéola est bordée par une structure en bourrelet nommée clivus, d'une épaisseur de 410 μm . Ailleurs, l'épaisseur moyenne de la rétine varie entre 180 μm à l'équateur et va en diminuant sur la périphérie (jusqu'à 100 μm). Ces différentes structures sont visibles par des coupes histologiques ou in vivo par l'imagerie de fond d'œil et de la Tomographie par Cohérence Optique (*Optical Coherence Tomography*) (OCT). Cette brève description de la rétine va nous permettre d'introduire les deux grandes pathologies pouvant l'affecter et les manifestations cliniques que celles-ci déclenchent. Mais auparavant, nous ouvrons une brève parenthèse sur les modalités permettant d'imager la rétine.

2.2 Modalités d'acquisition de l'imagerie rétinienne

La photoréception se faisant dans la rétine, celle-ci possède intrinsèquement une propriété particulièrement propice à l'imagerie : en effet, si les rayons lumineux se focalisent en son sein, il est possible de leur faire parcourir le chemin inverse pour reconstruire une image de celle-ci. L'observation de la rétine peut donc se faire de manière non invasive (c'est-à-dire sans injection d'agent de contraste, sans radiation ni opération chirurgicale).

De nombreuses modalités ont été développées pour la visualiser, cette section présente les deux principales que nous avons manipulées lors de cette thèse :

- L'imagerie de fond d'œil, qui consiste en une photographie en couleurs de la surface rétinienne. Une acquisition de fond d'œil est peu coûteuse, rapide à obtenir et est à haute résolution. Ces propriétés la rendent très populaire dans le cadre de programme de dépistage à large échelle.
- La Tomographie par Cohérence Optique (*Optical Coherence Tomography*) (OCT). Cette forme d'imagerie repose sur le principe de l'interférométrie et permet d'obtenir une estimation de la profondeur des différentes interfaces rétinienne selon un axe : cette représentation 1D est appelée A-scan. En balayant cet axe le long d'une ligne, on obtient une image 2D appelée B-scan comparable à une vue en coupe de la rétine. En déplaçant l'axe de coupe, on obtient une succession de B-scans formant un volume : l'OCT permet ainsi d'obtenir une représentation tridimensionnelle de la rétine.

2.2.1 Le fond d'œil

L'imagerie de fond d'œil est une appellation générique pour désigner une acquisition bidimensionnelle de la rétine par mesure de la quantité de lumière réfléchie à sa surface. En réalité, le terme de caméra fond d'œil (ou simplement fond d'œil) n'est qu'un dérivé de l'ophtalmoscope, où l'observation directe par le clinicien est remplacée par une mesure par un capteur photosensible. Dans son principe, l'ophtalmoscope est relativement simple : il focalise un faisceau lumineux incident qui en se réfléchissant permet d'observer la surface interne de l'œil. Dès 1851, Hermann von Helmholtz observait ainsi, à la lumière de sa bougie, la surface d'une rétine. L'ajout d'un capteur photographique (d'abord analogique puis numérique) a donné naissance aux premières caméras fond d'œil. Mais derrière cette appellation sont regroupées plusieurs technologies. Dans sa version la plus commune, le fond d'œil est un appareil photographique classique acquérant des images à une (ou trois pour la couleur) longueur d'onde donnée. La première caméra fond d'œil fut commercialement distribuée par Carl Zeiss en 1926. Depuis, les progrès techniques ont amené à proposer des améliorations en termes de champs de vue (*field of view*), de portabilité ou sur la possibilité de se passer d'un agent mydriatique pour dilater la pupille du patient. Il existe des dizaines de modèles commercialement disponibles, chacun possédant ses caractéristiques propres. La plupart des caméras sont dorénavant non mydriatiques et fonctionnent par mesure de quantité de lumière blanche réfléchie. Cette technologie offre l'avantage de fournir des images aux couleurs naturelles, c'est-à-dire comparables à celles observables à l'ophtalmoscope, à l'inverse d'une technologie basée sur une longueur d'onde unique. Aujourd'hui, les caméras disponibles sur le marché ont un champs de vue aux alentours de 45° et embarquent généralement un ensemble de tech-

nologies sophistiquées (post-traitements numériques, acquisitions multiples voire même pour la *JedMed Horus Scope* la capacité d'imager d'autres parties du corps (oreilles, nez, gorge). La résolution d'acquisition varie grandement, allant de 2 Mpixels pour la *VersaCam*, la *Kowa Genesis-D* ou encore la *OCULUS ImageCam 2* jusqu'à 12 Mpixels pour la *Kowa Nonmyd7*. Depuis peu, il existe également des modèles capables de faire des acquisitions à très large champ de vision, comme le propose la *California ultra-widefield retinal imaging*, avec un champ de vision de 200°. En 2016, une étude est publiée par Panwar et al. [8] recensant les principales caméras existantes et leurs caractéristiques. À noter qu'au moins trois versions de caméra fond d'œil se greffant à un téléphone portable existent désormais sur le marché.

À l'existence de ces différentes technologies d'acquisition, il faut ajouter un certain nombre de dérivés du fond d'œil classique ; l'annexe A en propose une brève description.

Adoption clinique de la fundoscopie

Malgré l'apparente simplicité de l'observation de la rétine par ophtalmoscopie, il est intéressant de noter que son usage est loin d'être simple ni universellement répandu en dehors de la communauté ophtalmologique. Par exemple, alors même que le *Consortium of Neurology Clerkship Directors* [9], rattaché à l'*American Academy of Neurology*, préconise la généralisation de l'examen de la rétine, celui-ci reste bien trop peu pratiqué lors du diagnostic de pathologies neurologiques. Une étude de 2008 menée auprès de 91 patients admis aux services d'urgences pour migraines soudaines et sévères montrait que seulement 42 d'entre eux (48%) avaient fait l'objet d'un examen de fond d'œil. Celui-ci aurait permis d'examiner leur nerf optique et de montrer potentiellement un œdème papillaire, une occlusion vasculaire ou l'absence de la pulsation des veines. Cette étude est renforcée par une seconde étude de 2013 par Thulasi et al. [10], menée auprès de 228 patients. Seuls 28 patients (12%) avaient reçu un examen par ophtalmoscope ; et plus problématique encore, aucun des 8,5% fond d'œil observés présentant des anomalies n'avait été détecté par les services des urgences. Faisant ce constat, un rapport de 2017 paru au sein du *American Academy of Neurology* [11], dresse l'inventaire des barrières entravant l'adoption de l'examen du fond d'œil en neurologie. Les conclusions obtenues décrivent un examen à l'ophtalmoscope trop contraignant dans les services d'urgence, du fait de facteurs multiples : la nécessité d'opérer à faible luminosité, l'emploi d'un agent mydriatique (dilatant la pupille) dont les effets secondaires sont parfois redoutés, un temps d'attente (d'action de l'agent mydriatique) trop long (30min)... Mais citant les résultats très encourageants de l'étude FOTO-ED [12–15], le rapport préconise l'adoption de l'examen fond d'œil via des technologies non mydriatiques, y compris embarquées au sein d'appareils d'usage commun (téléphone portable intelligent ou tablette). Le recours à la télémédecine est

également recommandé, séparant la prise de décision diagnostique de l'observation.

Sans anticiper sur la suite du présent document, il est intéressant de noter que ces recommandations s'accordent avec la problématique identifiée pour cette thèse, en particulier la nécessité d'avoir des systèmes capables de détecter des pathologies qui n'étaient pas nécessairement celles anticipées lors de l'acquisition.

2.2.2 Tomographie par Cohérence Optique (OCT)

L'OCT permet l'estimation de la profondeur des couches rétinienne. À l'inverse du fond d'œil, ce n'est pas une intensité lumineuse qui est mesurée, mais un temps de parcours de la lumière. Ce temps étant évidemment infinitésimal, il ne peut être mesuré directement. En revanche, une estimation précise peut s'obtenir par interférométrie. Son principe consiste à envoyer un faisceau de lumière au travers d'un miroir semi-réfléchissant qui séparera le faisceau en deux faisceaux cohérents. Le premier sera envoyé vers l'œil et le second vers un miroir amovible. Une partie de la lumière envoyée vers l'œil est réfléchi par chacune des couches rétinienne qu'elle rencontre (du fait d'une variation de l'indice de réfraction à chaque interface entre deux couches). On mesure la corrélation (ou plus exactement l'autocorrélation) entre le faisceau réfléchi par le miroir amovible (appelé référence) et celui réfléchi par une interface dans la rétine (appelé signal). Le pic d'autocorrélation est obtenu lorsque les deux faisceaux ont effectué le même trajet. En déplaçant le miroir amovible, on peut mesurer différents pics d'autocorrélation correspondant aux différents rayons réfléchis par les interfaces entre deux couches rétinienne et on obtient ainsi la profondeur de chacune des couches. La mesure de corrélation n'est cependant pas directe et passe par l'utilisation de l'interférométrie. En particulier, on peut montrer (Yaqoob et al. [16]) que pour une lumière de largeur de bande spectrale $\Delta\lambda$, les figures d'interférences mesurées obéissent à la relation :

$$P_{detection} = P_{ref} + P_{sig} + 2\sqrt{P_{sig}P_{ref}} \operatorname{sinc}\left(\frac{\pi(z_{ref} - z_{sig})}{l_c}\right) \cos(2\lambda_0(z_{ref} - z_{sig})) \quad (2.1)$$

où x_{ref}, x_{sig} correspondent respectivement à la distance parcourue par le faisceau de référence (resp. du signal), λ_0 à la longueur d'onde centrale du signal et $l_c = \frac{\pi}{\Delta\lambda}$ est appelée longueur de cohérence. La puissance détectée est maximale pour $z_{ref} = z_{sig}$. On peut donc bien mesurer ces pics en déplaçant le miroir. Pour obtenir des pics bien définis, on cherchera à diminuer la longueur de cohérence, d'où l'emploi d'une source d'éclairage à large bande spectrale (c'est à dire choisir un $\Delta\lambda$ grand).

L'imagerie de la profondeur (c'est-à-dire le long de l'axe z) en un point de coordonnées x, y est appelée A-scan. On obtient une représentation tridimensionnelle complète en balayant le

faisceau lumineux sur l'ensemble de la rétine. Une tranche (axe z-y ou z-x) de cette représentation est appelée B-scan, qui correspond donc à une succession de A-scans adjacents. Le premier A-scan obtenu avec cette technologie remonte au milieu des années 80 (Fercher et al. [17]), suivi du premier *B-scan* en 1991 (Huang et al. [18]) et la première acquisition in vivo de la rétine humaine en 1993 (Fercher et al. [19]). Depuis, l'OCT est devenu un instrument largement déployé en clinique et dont les performances n'ont cessé d'être améliorées. En 2008, une revue de littérature sur le sujet par Drexler et al. [20] identifiait les paramètres suivants comme critère de performance :

- La résolution axiale (profondeur ou axe z).
- La résolution transverse.
- Le temps d'acquisition.
- La sensibilité à la détection (c'est-à-dire la capacité à obtenir des acquisitions de bonne qualité y compris chez les patients atteints de pathologies).
- La profondeur de pénétration du faisceau lumineux.
- Le contraste de l'image finale.

En ce qui concerne la technologie du miroir amovible (appelée OCT dans le domaine temporel), le temps d'acquisition est dépendant du déplacement physique du miroir. Cette contrainte limite fortement le nombre d'acquisitions A-scans réalisables par seconde (aux alentours de 8000). Pour cette raison, des améliorations ont été proposées pour se défaire de la contrainte mécanique via la technologie de l'OCT fonctionnant dans le domaine spectral (ou fréquentiel). Pour expliquer simplement son fonctionnement, il faut revenir à l'équation fondamentale de l'interférométrie. Similairement à l'équation 2.1, on peut estimer la puissance reçue par le détecteur à une longueur d'onde donnée λ :

$$P_{detection}(\lambda) = P_{ref}(\lambda) + P_{sig}(\lambda) + 2\sqrt{P_{sig}(\lambda)P_{ref}(\lambda)} \cos(2\lambda(z_{ref} - z_{sig})) \quad (2.2)$$

On constate que la période de pulsation du terme $\cos(2\lambda(x_{ref} - x_{sig}))$ est proportionnelle à la distance $x_{ref} - x_{sig}$. En effectuant la transformée de Fourier du signal, on peut par conséquent mesurer les différentes fréquences (spatiales) contenues dans le signal, correspondant aux différentes valeurs de $z_{ref} - z_{sig}$ (seul z_{sig} variant en fonction des différentes profondeurs de couches). Deux approches existent pour effectuer la mesure du signal décrit dans l'équation 2.2. La première consiste à utiliser une source lumineuse large bande qui est décomposée au moment de la mesure en utilisant un spectromètre. Cet instrument est généralement composé d'un réseau de diffraction et de capteurs CCD (*charge coupled device*). Le réseau de diffraction permet de séparer le chemin de la lumière en fonction de sa longueur d'onde. Les différents chemins seront mesurés en plusieurs points grâce aux capteurs CCD positionnés

adéquatement. On obtient alors directement un motif d'interférences ressemblant à celui du domaine temporel (mais dépendant de λ), chaque pixel correspondant à une longueur d'onde. La seconde méthode employée est d'utiliser une source à bande spectrale étroite modulable (appelée source à balayage). Le signal de l'équation 2.2 est donc généré pour différentes longueurs d'onde modulées (c'est-à-dire espacées linéairement) et discrétisées $\lambda(t) = \lambda_0 + \delta\lambda t$. Comme pour l'approche par spectromètre, la transformée de Fourier du signal permet d'identifier les différents pics correspondants aux différentes valeurs de $z_{ref} - z_{sig}$. Mais à la différence de cette dernière, l'acquisition ne requiert qu'un seul point de mesure (la mesure étant faite au cours du temps) et non d'un réseau de diffraction. Bien que fondamentalement similaire à la première méthode, la technologie à balayage est moins sensible aux bruits (inhérents à une source large bande). Néanmoins, elle introduit des défis spécifiques à la construction d'une source lumineuse modulable. C'est ainsi que jusqu'en 2008, toutes les sources fonctionnaient dans la bande 1300 nm [20], là où une bande centrée sur 1040 nm serait plus adaptée à une pénétration dans les tissus choroïdiens.

Jusqu'ici, les approches décrites ne font l'acquisition qu'en un seul point (x, y) à la fois. Mais plusieurs études ont été proposées des acquisitions transverses parallèles (sous le nom de Tomographie *En face*) [21–23]. Ces travaux reposent sur l'amélioration continue des capteurs numériques, capables désormais d'acquérir des images 2D. Ces méthodes sont intrinsèquement plus rapides et moins sensibles au mouvement parasite de l'œil. Elles ouvrent la voie à l'acquisition 3D, vidéo et en temps réel.

En ce qui concerne la résolution des volumes acquis, la tomographie présente la particularité d'avoir une résolution axiale indépendante des résolutions transverses. Ainsi, si la première est dépendante de la longueur de cohérence de la source, la résolution transverse est liée à une problématique classique en optique : la focalisation. En particulier, au passage dans la pupille, la lumière est diffractée, entraînant un phénomène d'aberrations chromatiques. Pour faire face à ce défi, l'intégration de lentilles adaptatives (*Adaptive Optics*) a ouvert la voie à une imagerie à l'échelle cellulaire.

Tomographie par effet Doppler Dès la fin des années 90, le principe de l'OCT a été étendu à l'imagerie fonctionnelle [24]. En particulier, l'application du principe de Doppler a permis d'étudier le flux sanguin (Leitgeb et al. [25, 26]). On utilise le changement de phase induit dans les rayons réfléchis du fait du mouvement du sang. Ces travaux ont montré la capacité théorique de l'OCT (dans le domaine spectral, car on y étudie la phase du signal) à estimer la pression artérielle et veineuse quantitativement à l'échelle d'un vaisseau individuel. Cependant, l'évaluation quantitative du flux sanguin par effet Doppler nécessite de connaître l'angle d'incidence du faisceau lumineux avec le vaisseau et donc une modélisation en amont

de la structure vasculaire, ce qui en fait une méthode complexe malgré les attentes qu'elle suscite. À l'échelle clinique, ces défis importants ont limité l'utilisation de la tomographie par effet Doppler à la reconstruction 3D de l'arbre vasculaire (acquisition appelée « Angiographie par Cohérence Optique », ou OCT-A).

Variantes et dérivées de l'OCT Il existe d'autres dérivées de l'OCT qui ne seront pas abordées dans le présent document : on peut citer des applications reposant sur les propriétés biréfringentes de certains tissus pour les caractériser, ou encore le *Enhanced Depth Imaging Optical Coherence Tomography (EDI-OCT)* qui repose sur les propriétés de symétrie de la transformée de Fourier inverse (d'un signal réel) pour améliorer les contrastes sur un volume acquis par une machine OCT spectrale classique.

2.3 Pathologies rétinienne et contexte clinique

Il existe un très grand nombre de maladies rétinienne, aussi bien susceptibles de n'affecter que la rétine que d'être le symptôme d'une pathologie systémique. Une énumération exhaustive de celles-ci n'est pas le propos de cette section, nous nous contenterons d'une synthèse sur les deux principales maladies que sont la rétinopathie diabétique et la dégénérescence maculaire liée à l'âge. Le choix de ces maladies est dicté par deux considérations :

1. Leur prévalence et incidence respectives, qui en font des maladies de première importance en termes de santé publique.
2. En conséquence, l'effort déployé ou ambitionné pour aboutir à des campagnes de dépistage à plus ou moins large échelle. Pour arriver à cette fin, l'intégration des nouvelles technologies (que ce soit d'acquisition, de télémedecine ou d'analyse automatique) fait consensus. Cela justifie donc d'y prêter une attention toute particulière.

2.3.1 La rétinopathie diabétique

La RD est la principale cause de déficience visuelle au sein de la population adulte en âge de travailler. Elle se développe en conséquence d'une complication répandue du diabète sucré. Ce dernier est lui-même causé par une défaillance des mécanismes de régulation de la glycémie entraînant une hyperglycémie. Or, un taux élevé de glucose dans le sang occasionne des dommages sur les petits et gros vaisseaux sanguins. La cible principale du glucose est l'endothélium (tissu constituant la paroi interne des vaisseaux, formée de cellules épithéliales) et la lame basale (sur laquelle reposent les cellules épithéliales et chargée des apports nutritifs à l'endothélium). Un taux élevé de glucose et d'insuline dans le sang peut avoir pour consé-

quence de modifier les réactions des cellules de l'endothélium et d'augmenter la viscosité du sang, modifiant ainsi chimiquement certaines protéines. Dans le cas de la rétine, ces effets peuvent entraîner deux types de réactions :

- Une ischémie, que l'on définit comme la diminution de l'apport sanguin artériel à un organe, généralement du fait d'une obstruction de l'artère. Dans la rétine, cette obstruction peut conduire à la croissance de néovaisseaux pouvant entraîner des saignements et causer un détachement rétinien. Si ce stade est atteint, on parle alors de RD proliférative.
- La rupture de la barrière hématorétinienne. Le liquide exsudé peut provoquer un enfllement de la macula et endommager les photorécepteurs. Cette complication a pour nom un œdème maculaire diabétique (OMD).

L'ischémie peut causer un blocage du transport axonal (déplacement de mitochondries, lipides, protéines et autres cellules dans le corps du neurone, permettant le renouvellement de celui-ci). L'accumulation axoplasmique dans la couche des fibres optiques entraîne des lésions appelées nodules cotonneux, d'apparence blanche et aux frontières floues.

Cliniquement, les premiers signes de RD sont l'apparition de microanévrismes et d'hémorragies intrarétinales. À un stade plus avancé de la maladie, la rétine s'épaissit (formation d'un œdème) et des exsudats (dépôts de lipides) apparaissent pouvant causer une perte d'acuité visuelle. Si la progression de la maladie se poursuit, elle peut atteindre le stade prolifératif, dont le nom provient de l'apparition de nouveaux vaisseaux dans le disque optique, la rétine ou l'iris. Ces néovaisseaux ont pour conséquence potentielle un détachement rétinien ou un glaucome néovasculaire. C'est à ce stade que le risque de perte totale de la vision est le plus important.

Prévalence de la maladie En étant une conséquence du diabète et en particulier celui de type 2, la prévalence de la RD est directement liée à celle du diabète. Aux États-Unis par exemple, on estime à près de 11% la part de la population âgée de plus de 25 ans atteinte de diabète, même si on peut noter que cette proportion varie fortement en fonction de l'ethnicité, allant de 12,6% pour les populations afro-américaines à 11,8% chez les hispaniques et 7,0% chez les européennes. L'accélération de l'incidence de l'obésité infantile fait craindre à une expansion croissante de la maladie. Par ailleurs, la maladie ne se restreint pas aux frontières des États-Unis et touche tous les bassins régionaux du monde à tel point qu'elle a été qualifiée de pandémie dès la fin des années 90 par des organismes de santé publiques transnationaux (par exemple la *European Society of Cardiology* [27] ou l'Organisation mondiale de la santé [28]). Logiquement, la prévalence de la RD suit cette tendance mondiale et on dénombre par conséquent une part croissante de la population mondiale atteinte de la maladie. En 2016, on

estime à près de 93 millions le nombre d'individus atteints de RD, parmi lesquels se trouvent 28 millions de personnes pour lesquelles la maladie menace directement la vision.

Progression de la maladie La sévérité de la RD est souvent évaluée en se basant sur une échelle prédéfinissant les étapes de la progression. Il existe plusieurs systèmes d'échelles, dont le plus connu est l'*ETDRS* [29]. Une reproduction des étapes de progression est proposée dans le tableau 2.1. Il est intéressant de noter que les symptômes définissant ces grades sont tous visibles à l'imagerie de fond d'œil.

2.3.2 La dégénérescence maculaire liée à l'âge

Comme son nom l'indique, la DMLA est une maladie affectant la macula. Elle est la principale cause de cécité chez les personnes âgées de plus de 50 ans dans les pays développés. Comme la DMLA se concentre dans la région maculaire, cette cécité est rarement complète, car elle n'affecte pas la vision périphérique. Les causes de la maladie restent mal connues, mais sont suspectées d'être polygéniques, incluant des facteurs de risques environnementaux (tabagisme, surpoids). Il existe deux formes de DMLA : la première, dite sèche, correspond à une atrophie de la macula et représente la très vaste majorité des cas. La seconde, dite exsudative ou humide, correspond à l'apparition de néovascularisation choroïdienne. Bien que plus rare, elle évolue beaucoup plus rapidement que la première. La majeure partie des symptômes que nous synthétisons ici provient des travaux de Kolar et. [30].

La DMLA se caractérise par l'un (ou plusieurs) des symptômes suivants :

- La présence de drusens de taille intermédiaire. Un drusen est une lésion d'apparence jaune plus ou moins contrastée. Histologiquement, ils sont situés au niveau de l'interface avec l'épithélium pigmentaire (*Retinal Pigment Epithelium*) (RPE). Une taille intermédiaire correspond à une taille comprise entre $63\mu\text{m}$ et $125\mu\text{m}$.
- Des anomalies de la RPE, en particulier de l'hypopigmentation ou de l'hyperpigmentation. L'hypopigmentation est associée à un amincissement de la RPE et en conséquence à un plus faible taux de mélanine. L'hyperpigmentation quant à elle est une conséquence d'une migration ou d'une prolifération des cellules de la RPE vers les couches supérieures (dans la direction de la surface de la rétine)
- La présence de pseudodrusen réticulaire. Il s'agit de dépôts de matière situés dans les couches supérieures à la RPE.
- La présence d'un ou plusieurs des marqueurs suivants : une atrophie géographique de la RPE -qui laisse apparaître la choroïde voire même de la sclère à l'imagerie de fond d'œil-, une néovascularisation choroïdale, des traces de prolifération angiomateuse

TABLEAU 2.1 Échelle de sévérité de la Rétinopathie Diabétique (RD) telle que définie par l'étude *ETDRS*

Grade	Degré de sévérité	Indices cliniques visibles sur le fond d'œil
0	Pas de RD apparente	Pas d'anomalies
1	RD non-proliférative légère	Microanévrismes
2	RD non-proliférative modérée	Microanévrismes + autres lésions mais pas les marqueurs de la RD sévère
3	RD non-proliférative sévère Définition américaine	Règle des 4-2-1 Signes parmi les suivants et pas de signe de RD proliférative : • Hémorragies intra-rétiniennes sévères et microanévrismes dans chacun des quatre quadrants • Veines boudinés dans au moins deux quadrants • Anomalie microvasculaire intra-rétinienne modérée dans au moins un quadrant
	Définition internationale	Signes parmi les suivants et pas de signe de RD proliférative : • Plus de 20 hémorragies intra-rétiniennes dans chacun des quatre quadrants • Veines boudinés dans au moins deux quadrants • Importante anomalie microvasculaire intra-rétinienne dans au moins un quadrant
4	RD proliférative	Parmi les suivants : • Néo-vascularisation • Hémorragie pré-rétinienne et/ou intra-vitréenne

rétinienne (*Retinal Angiomatous Proliferation*) -en simplifié, une néovascularisation démarrant depuis la rétine.

Prévalence de la maladie En 2004, la DMLA affectait 1.75 millions de personnes âgées de 40 ans ou plus aux États-Unis à un stade relativement avancé, c’est-à-dire avec des signes de néovascularisation ou d’atrophie géographique dans au moins un œil. À des stades plus précoces de la maladie, 7.3 millions de personnes présentent de larges drusens. Là encore, Wong et al. [31] font état d’une variabilité importante au sein des différentes ethnicités. En 2020, les estimations portent à 196 millions le nombre de personnes atteintes avec des projections pour 2040 atteignant les 288 millions.

Progression de la maladie S’il existe plusieurs échelles de gradation de la maladie, nous reproduisons ici la classification proposée en 2013 par Ferris et al. [32] (Tableau 2.2). Ce système est lui-même dérivé des conclusions de l’étude de référence AREDS *Age-Related Eye Disease Study* menée entre 1992 et 2006.

TABLEAU 2.2 Echelle de sévérité de la DMLA

Degré de Sévérité de la DMLA	Indices cliniques
Pas d’effets liés au vieillissement apparents	Pas de drusen et pas d’anomalies pigmentaires
Vieillessement normal	Seulement des petits drusens et pas d’anomalies pigmentaires
DMLA précoce	Présence de drusens de taille moyenne et pas d’anomalies de la RPE.
DMLA intermédiaire	Au moins une des caractéristiques suivantes : — De nombreux drusens de taille moyenne — Au moins un drusen de grande taille ($> 125 \mu\text{m}$) — Une atrophie géographique n’affectant pas la fovéa.
DMLA avancée (néo-vasculaire, humide ou exsudative)	Au moins une des caractéristiques suivantes : — Une atrophie géographique affectant la fovéa. — Une néovascularisation au niveau de la macula, dont il existe différents types (Neovascularisation choroïdale (<i>Choroidal neovascularization</i>) (CNV), détachement hémorragique de la RPE, exsudats durs...)

CHAPITRE 3 REVUE DE LITTÉRATURE

La recherche informatique appliquée à l'imagerie médicale se trouve à l'intersection de deux domaines et a par conséquent deux objets d'études : la conception d'algorithmes novateurs et la validation de ceux-ci suivant un protocole simulant un déploiement clinique. Les fondations de ce champs de recherche reposent donc à la fois sur des théories mathématiques statistiques et sur des considérations plus empiriques portant sur l'efficacité et l'utilisabilité des algorithmes dans un contexte clinique. Pour introduire notre recherche, on ne peut faire abstraction d'une description des emprunts faits à l'apprentissage machine et aux statistiques en général, emprunts réalisés non sans de nécessaires (et parfois substantielles) adaptations. Afin de faciliter la compréhension du lecteur non initié, cette revue de littérature s'ouvre sur une introduction des principes théoriques constituant les briques fondamentales de la reconnaissance automatique d'images par apprentissage profond. Ainsi, les sections allant de 3.1 à 3.1.5 ont essentiellement une visée pédagogique. Les algorithmes qui y sont mentionnés appartiennent à la trousse à outils du chercheur, leur application n'est pas spécifique à un domaine en particulier. Nous y couvrons les opérateurs fondamentaux sur lesquels se basent les réseaux de neurones, un survol historique des architectures qui ont marqué ces dix dernières années (dont plusieurs réapparaîtront dans nos expériences) et un rappel sur les méthodes d'entraînement de celles-ci.

À partir de la section 3.2.2, nous concentrons notre revue de littérature sur les éléments qui occupent le cœur de cette thèse : ainsi, dans la section 3.2, nous introduirons le concept de généralisation et les éléments statistiques qui appartiennent à ce domaine. Cette présentation est découpée en une première introduction théorique suivie de considérations plus empiriques sur les questions d'adaptation de domaines et d'apprentissage sur cible bruitée.

La section 3.3 présente les notions d'explicabilité et d'interprétabilité des modèles, c'est-à-dire la compréhension qu'un utilisateur humain peut avoir de leur fonctionnement. Plus spécifiquement, nous nous concentrons sur le champs de recherche dédié aux méthodes de génération d'attribution locale.

Enfin, à partir de la section 3.5, nous aborderons plus spécifiquement les applications propres à l'ophtalmologie des modèles d'apprentissage machine, en discutant de la segmentation sémantique des lésions du fond d'œil et des approches locale et globale de classification de maladies.

3.1 Réseaux de neurones dédiés à la vision par ordinateur

L'apprentissage machine est un champs englobant de nombreux concepts et algorithmes inventés au fil du développement de l'informatique. S'il ne fallait leur trouver qu'un seul point commun, ce serait de s'appuyer sur un ensemble de données observées pour construire une représentation statistique des relations qui régissent ces données. En cela, l'apprentissage machine se distingue des modélisations mathématiques ou physiques plus traditionnelles car il ne cherche pas à construire une théorie identifiant explicitement les mécanismes de cause à effet ; mais plutôt à exploiter les corrélations statistiques qui expliquent au mieux les données observées. L'hypothèse qui prévaut est que si l'observation est suffisamment exhaustive, le modèle statistique s'approchera au plus près de la loi réelle qui régit le phénomène observé. Dans ce vaste univers qu'est l'apprentissage machine, nous n'aborderons directement qu'une sous-catégorie, dite de l'apprentissage profond. Celle-ci a largement accaparé les devants de la scène scientifique mais aussi médiatique. Ses contours ne sont pas forcément univoquement définis, mais la grande majorité des travaux qui se revendique de l'apprentissage profond repose sur une forme plus ou moins évoluée de réseaux de neurones. Afin d'introduire ce domaine, nous souhaitons définir en premier lieu les grandes briques fonctionnelles utilisées par les réseaux de neurones. Les deux prochaines sous-sections sont donc à aborder comme une sorte de lexique simplifié ; ce n'est qu'à partir de la section 3.1.1 que nous rentrerons dans les détails de la littérature sur le sujet.

Le neurone formel

Dans sa conception originale, le neurone formel s'inspire de son équivalent biologique : ce dernier possède plusieurs dendrites (vues comme des entrées du système) et un axone (représenté comme sa sortie). Les dendrites reçoivent un influx nerveux en provenance d'autres neurones (plus exactement de leur terminaison axonale) au niveau des synapses : ces influx sont modélisés sous forme de signaux numériques. L'opération réalisée par le neurone est modélisée comme une somme pondérée de ses entrées, puis, si cette somme franchit un certain seuil, sa transmission à sa sortie. Ainsi, dans sa forme la plus épurée, le neurone formel à N entrées réalise simplement l'opération suivante :

$$y = \sigma\left(\sum_{i=1}^N x_i w_i + b\right) \quad (3.1)$$

où $\mathbf{x} \in \mathbb{R}^N$ est son entrée et $y \in \mathbb{R}$ sa sortie. L'opérateur σ représente une fonction de seuil, aussi dite d'activation. Elle est non-linéaire et distingue le neurone formel d'une simple

projection linéaire. Les paramètres $\mathbf{w} \in \mathbb{R}^N$ et $b \in \mathbb{R}$ (appelés respectivement poids et biais) sont ses variables d'ajustement qui, modifiées en fonction de l'objectif visé, permettent au neurone de produire la sortie attendue et donc « d'apprendre ». L'équation 3.1 peut se réécrire de manière équivalente sous sa forme vectorielle : $y = \mathbf{w}^T \mathbf{x} + b$.

En pratique, on n'utilise jamais un seul neurone isolé mais plutôt une couche de neurones (qui n'est rien d'autres que le calcul parallèle de M neurones). Avec $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$, $\mathbf{b} \in \mathbb{R}^M$ et $\mathbf{W} \in \mathbb{R}^{N \times M}$, l'équation d'une couche neuronale s'obtient de manière similaire à celle de l'équation 3.1 :

$$\mathbf{y} = \sigma(\mathbf{W}^T \mathbf{x} + \mathbf{b}) \quad (3.2)$$

Enfin, revenons sur la fonction d'activation σ : de nombreuses propositions existent dans la littérature, nous en recensons quelques unes sur la figure 3.1. Elles modélisent en général un effet de saturation (limite finie vers $+\infty$ / $-\infty$), qui s'interprète comme un seuil au delà duquel le neurone s'active.

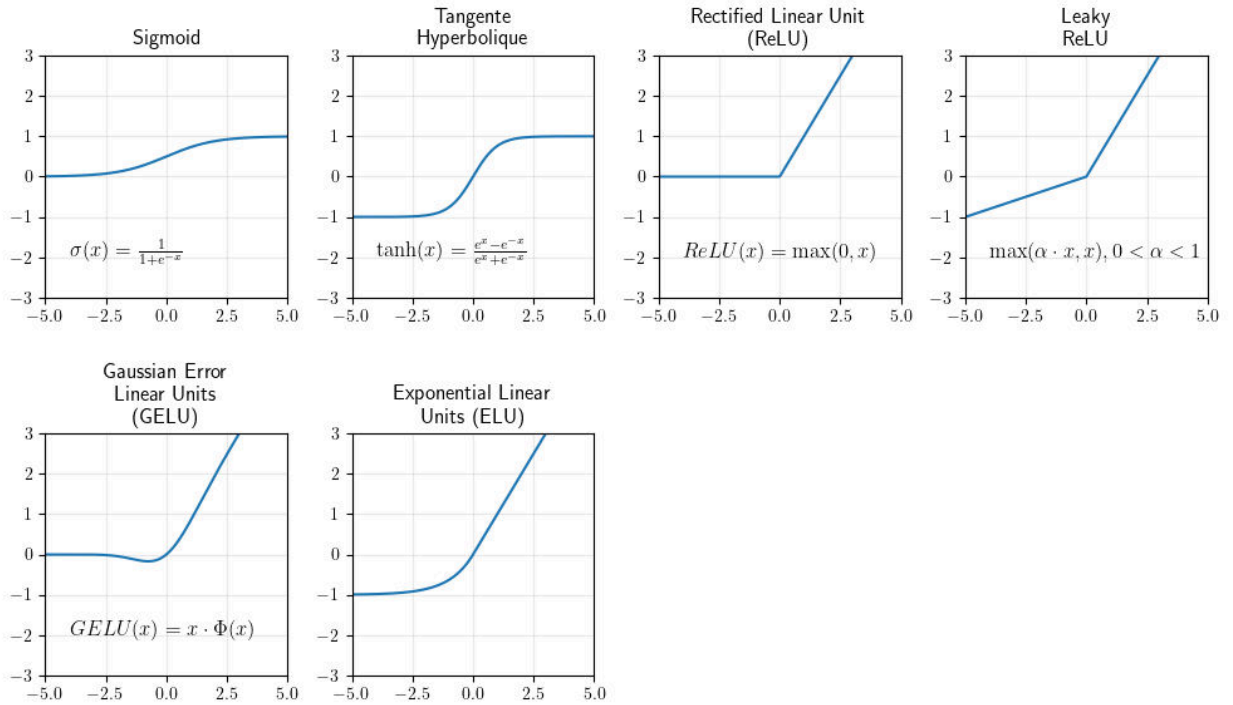


FIGURE 3.1 Fonctions non-linéaires usuelles placées en sortie de neurones

Le réseau de neurones multicouches

En ayant introduit le neurone formel et son corollaire, la couche de neurones, nous disposons des bases nécessaires pour définir le premier réseau de neurones, dit multicouches. Son principe est simple : à l’instar du cerveau et son imbrication complexe de neurones, il est possible d’organiser un réseau artificiel sous la forme d’une succession de couches, connectées les unes aux autres de manière séquentielle. Autrement dit, la sortie de la couche l est fournie en entrée de la couche $l + 1$. Pour chacune, l’équation 3.2 est respectée, même si chaque couche peut disposer d’une fonction d’activation qui lui est propre. Ce faisant, un réseau peut s’interpréter comme une composition (au sens mathématique) alternant projection linéaire et non-linéarité via la fonction d’activation :

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) = (\sigma^{(L)} \circ f^{(L)}) \circ \dots \circ (\sigma^{(l)} \circ f^{(l)}) \circ \dots \circ (\sigma^{(0)} \circ f^{(0)})(\mathbf{x}) \quad (3.3)$$

Sémantiquement, toutes les couches d’indice $l \in \{0, 1, \dots, L - 1\}$ sont appelées couches cachées dont la quantité L est référée sous le terme de profondeur du réseau. Par analogie, on parle de largeur pour désigner le nombre de neurones par couche. Ces deux paramètres (largeur et profondeur) font partie des choix architecturaux à faire dans la conception d’un modèle. Depuis les années 80, l’expressivité de $\mathcal{F}$ (c’est-à-dire sa capacité à approximer n’importe quelle fonction) est largement étudiée, dépendamment de la largeur et de la profondeur choisies. Les travaux de Cybenko G. sont précurseurs en la matière : il y introduit un des Théorèmes d’Approximation Universelle, démontrant la capacité d’un réseau de profondeur $d = 1$ (une couche cachée) utilisant la fonction sigmoïde à approcher n’importe quelle fonction dès lors que sa largeur est assez grande (possiblement infinie). Des versions plus modernes de ces travaux généralisent ce théorème à différents types de non-linéarité. Depuis les travaux de Zhou et al. [33] la tendance récente est à promouvoir une largeur bornée et une profondeur accrue. Cai et al. [34] explicitent, sous certaines conditions, une borne minimale à la largeur nécessaire.

Quelques opérateurs usuels

Dans la trousse à outils du concepteur de réseaux de neurones, d’autres opérateurs reviennent très fréquemment, sans que leur paternité ne soit très clairement définie : c’est en particulier le cas du *Pooling* et du *Softmax*¹.

1. Nous conservons dans cette thèse les termes anglais à défaut d’une traduction adoptée unanimement par la communauté francophone

Le Pooling est une opération de sous-échantillonnage très simple : au sein d’une fenêtre donnée, il s’agit d’agréger la sortie de plusieurs neurones en une seule valeur. En pratique, on utilise fréquemment deux types de pooling :

- Le *Max-Pooling* qui consiste à prendre la valeur maximale parmi les neurones dans la fenêtre considérée.
- Le *Mean-Pooling* qui effectue la moyenne des neurones dans cette fenêtre.

Il n’existe pas nécessairement de justification théorique sur l’utilisation de l’une ou l’autre des méthodes ; on peut simplement noter que le Max-Pooling est nettement prévalent.

Le Softmax est une opération de normalisation $f : \mathbb{R}^d \mapsto [0, 1]^d$ définie par :

$$f(\mathbf{x}) = \frac{\exp(\mathbf{x})}{\sum_i \exp(\mathbf{x}_i)} \quad (3.4)$$

Généralement appliqué en sortie de la dernière couche d’un réseau de neurone, le Softmax ramène toutes les valeurs des neurones dans un intervalle fermé $[0, 1]$. Par ailleurs, il contraint la somme de ces neurones à valoir 1. Pour cette raison, la sortie d’un softmax s’interprète généralement comme une distribution discrète de probabilités.

3.1.1 Réseaux de Neurones Convolutifs

La conceptualisation et l’étude de nouvelles architectures de réseaux de neurones ont constitué une majeure partie des publications récentes (période 2012-2023). Comme le formulent Alzubaidi et al., cet aspect de la recherche représente le noyau dur de l’*Apprentissage Profond* [35].

Nous mettons ici l’accent sur les principales avancées obtenues en reconnaissance automatisée dans le domaine de la vision par ordinateur, donc sur des images. Dans ce domaine et bien avant l’émergence des réseaux neuronaux, l’opérateur de convolution est reconnu comme une clé de voûte de bien des algorithmes de traitement d’images. Connue pour ses propriétés dans le domaine du filtrage, la convolution est réinterprétée par Le Cun et al. [36] qui proposent de rendre son noyau entraînable par descente de gradients. Cette idée, sans être conceptuellement révolutionnaire, a une grande importance théorique et pratique. La convolution n’étant qu’une projection linéaire locale, son interprétation sous le prisme de la théorie du neurone formel est assez directe. L’idée fondamentale des auteurs est de remplacer les premières couches de neurones classiques par des couches convolutives, décomposant ainsi le réseau en une première partie liée à l’extraction de caractéristiques locales et une seconde à l’aggrégation et la classification de celles-ci. L’architecture résultante, appelée LeNet-5, devient donc le premier CNN. Elle est composée de 4 couches convolutives suivies d’un Per-

ceptron Multicouches (PMC) à deux couches. Dans le détail, chaque couche convolutive est constituée d'un ensemble de filtres, dont les sorties sont concaténées pour former un tenseur de cartes de caractéristiques. Formellement, notons f^k la fonction réalisée par la couche convolutive k . Celle-ci est paramétrée par un vecteur de poids $\mathbf{W}^k \in \mathbb{R}^{I \times O \times M \times N}$ et de biais $\mathbf{B} \in \mathbb{R}^I$. Les dimensions $M \times N$ représentent la taille du noyau de convolution. Cette fonction reçoit un tenseur d'entrée $\mathbf{X}^k \in \mathbb{R}^{I \times H \times W}$ et produit en sortie un tenseur $\mathbf{Y}^k \in \mathbb{R}^{O \times H' \times W'}$

$$\mathbf{Y}^k = f^k(\mathbf{X}^k) \quad (3.5)$$

L'équation détaillée d'une couche convolutive est donnée par :

$$\mathbf{Y}^k[o, h, w] = \sum_{i=1}^I \sum_{m,n=1}^{M,N} \mathbf{X}^k[i, h + m - \lfloor M/2 \rfloor, w + n - \lfloor N/2 \rfloor] \cdot \mathbf{W}^k[i, o, m, n] + \mathbf{B}[i] \quad (3.6)$$

La première dimension I correspond aux canaux du tenseur d'entrée. Ainsi, $\mathbf{X}^0$ qui correspond à l'image d'entrée du réseau, possède généralement un ou trois canaux (soit monochrome soit « rouge, vert, bleu ») et $I \in \{1, 3\}$. Plusieurs propriétés se dégagent de cette définition de convolution neurale :

- Si $\mathbf{W}$ est symétrique, l'opération de convolution décrite correspond à la définition de la corrélation au sens du traitement du signal.
- f^k est linéaire. En conséquence, $f^{k'} \circ f^k$ est également linéaire ; par extension, un réseau constitué d'une succession de couches convolutives ainsi définies se réduirait à une application linéaire ; ce qui n'est pas souhaitable pour l'expressivité de l'architecture globale. Pour cette raison, on introduit une non-linéarité en sortie de chaque couche convolutive, à l'instar du neurone formel standard.
- La taille du masque de convolution $M \times N$ correspond à la notion de champs de vision de la couche, autrement dit du nombre de points de l'image d'entrée considéré pour le calcul de la sortie. En conséquence, la taille du masque est indicatrice du contexte spatial que la couche convolutive est en mesure de prendre en compte dans son calcul de sortie.
- L'opération de convolution est équivariante en translation, ce qui revient à dire que la même projection linéaire appliquée localement se fait sur toute l'image.

Dans la grande majorité des cas, les architectures se composent de successions séquentielles de couches de convolutions, réduisant progressivement la résolution spatiale des cartes de caractéristiques tout en accroissant la largeur de chaque couche (le nombre de cartes extraites). De la sorte, le réseau effectue une agrégation spatiale de l'information contenue dans l'image, tout en allouant une dimension croissante au vecteur caractéristique assigné à

chaque position spatiale, permettant à cette agrégation de gagner en abstraction (c'est-à-dire en capacité à représenter des structures de plus en plus complexes et spatialement éloignées dans l'image). Dans les dernières couches, la résolution de l'image d'entrée se retrouve suffisamment compressée pour qu'elle puisse facilement se représenter sous la forme d'un vecteur, qu'il est possible de classifier via une succession de couches de neurones totalement connectés. L'intérêt de l'approche est d'automatiser entièrement l'étape d'extraction et de sélection de caractéristiques, la partie convolutive du réseau étant entraînée à la faire conjointement avec la classification. Comme nous le verrons, l'apprentissage présente l'avantage théorique de garantir l'extraction de caractéristiques localement optimales.

Khan et al. [37] proposent une revue de littérature détaillée des avancées techniques et conceptuelles dans le domaine des CNNs au cours de la dernière décennie. Nous en dressons ici une synthèse présentant les grandes architectures représentant des jalons importants dans l'histoire des réseaux de neurones. Au delà du seul intérêt historique, cette présentation introduit les éléments architecturaux utilisés par nos propres modèles. La modularité est en effet une notion fondamentale pour comprendre le succès des réseaux de neurones modernes.

AlexNet Proposée par Krizhevsky et al. [38], AlexNet (du nom de son auteur) est l'une des plus emblématiques architectures de CNN. Quinze ans après les travaux précurseurs de Le Cun et al. [36], AlexNet surpasse spectaculairement l'état de l'art dans la classification de grande base de données d'images (en l'occurrence la version de 2010 du ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [39], constituée de 1.4 millions d'images réparties en 1000 classes). Outre une optimisation des calculs par parallélisation sur Processeur Graphique (GPU), AlexNet présente trois évolutions notables (par ailleurs également introduites en collaboration avec Hinton et al. [40]) :

- L'utilisation d'une normalisation de la réponse locale en sortie de certaines couches.
- La régularisation du modèle pour éviter le sur-apprentissage par utilisation de la technique du *Dropout*, c'est-à-dire la mise à zéro aléatoire de certains neurones.
- L'utilisation de la fonction *Rectified Linear Units* (RELU) en sortie des couches convolutives (voir Figure 3.1)

Ce modèle marque le début du regain d'intérêt pour les CNNs et une prise en compte de l'intérêt des GPU pour les calculs hautement parallélisables (et autres que les opérations graphiques qui leur étaient réservées). En 2007, NVIDIA, fabricant leader sur le marché des GPUs met à disposition *Compute Unified Device Architecture* (CUDA), une librairie proposant une extension du langage C pour le calcul parallèle sur GPU. En 2014, prenant la mesure de l'importance que prennent les réseaux de neurones, l'entreprise sort *cuDNN*, une librairie additionnelle offrant un ensemble d'implémentations pré-écrites pour les opérations

usuelles qu'utilisent ces algorithmes.

VGG L'architecture VGG (du nom du laboratoire dont elle est issue, le *Visual Geometry Group*) [41] est également considérée comme une étape majeure dans l'évolution des CNNs. Structuellement, il s'agit toujours d'une succession de couches convolutives ; mais le VGG en accumule davantage, accroissant ainsi la profondeur du réseau. Pour y arriver, Simonyan et al. proposent de remplacer les masques de convolutions du AlexNet (originellement de taille 7×7 ou 5×5) par des masques plus petits (3×3 voire 1×1). Le nombre de paramètres entraînés étant égal au carré de la taille du masque, réduire cette dernière permet d'augmenter le nombre de couches. Le réseau gagne ainsi en profondeur. Cependant, un masque de taille 5×5 a un champs visuel réceptif naturellement plus élevé qu'un masque plus petit, ce qui peut sembler un argument important en faveur des masques de grande taille. Simonyan et al. le contrent en arguant que le champs réceptif s'accroît linéairement à chaque couche convolutive. Suivant cette logique, il est possible d'obtenir un champs réceptif de taille 5×5 en utilisant deux couches de taille 3×3 successives (la figure 3.2 illustre ce phénomène) : ce faisant, par rapport à un seul masque de 5×5 , on réalise une économie de $5 \times 5 - (2 \cdot (3 \times 3)) = 7$ paramètres (soit 28% de moins). De plus, à chaque opération de sous-échantillonnage insérée entre deux couches (on parle de *pooling*), le champs réceptif visuel théorique est doublé.

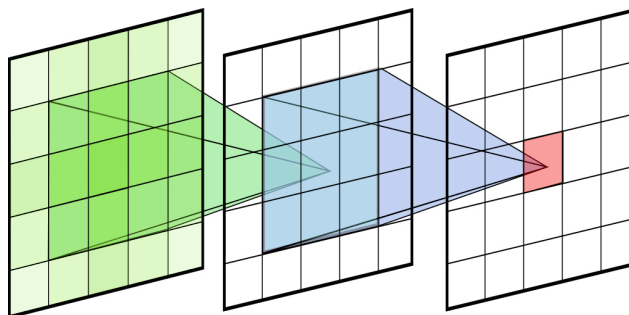


FIGURE 3.2 Illustration de la notion de champs visuel réceptif obtenu par la mise en séquence de deux couches convolutives de masque 3×3 . Pour obtenir la valeur du pixel rouge, il faut prendre les 3×3 pixels de l'image de la couche précédente, eux-même tous obtenus en utilisant les 5×5 pixels de la couche d'entrée. On peut donc parler de champs visuel réceptif de 5×5 pour la troisième couche.

Notons cependant que la notion de champs visuel théorique ne correspond pas à une réelle mesure d'importance d'un pixel d'entrée de réseau vis-à-vis d'un pixel en sortie d'une des couches convolutives. Il s'agit plutôt d'une borne supérieure, en pratique jamais atteinte, sur le contexte spatial pris en compte dans l'estimation de la sortie d'une couche donnée. Luo et al. proposent une estimation probabiliste du champs visuel réceptif effectif [42] faisant

l'hypothèse d'une distribution Gaussienne des noyaux des masques convolutifs. Ils démontrent que le champs visuel réceptif effectif est en pratique bien inférieur à son estimation théorique.

Inception-V1 Contemporain du VGG, Szegedy et al. proposent le GoogleNet, s'appuyant sur un nouveau type de module appelé *Inception-block* [1]. Les auteurs documentent également l'utilisation d'un certain nombre de techniques permettant de réduire le nombre de paramètres du réseau par couches (et donc d'augmenter le nombre de celles-ci). En particulier, leur architecture fait usage de convolution de taille 1×1 . En apparence, une telle opération n'est qu'une combinaison linéaire des cartes de caractéristiques qui lui sont fournies. Elle permet cependant d'en réduire la dimension, à l'instar d'un opérateur de compression. De plus, l'ajout d'une fonction non-linéaire à la sortie de chacune des convolutions 1×1 permet la composition de fonctions de plus en plus complexes. Quant à lui, le bloc *Inception* (Figure 3.3) a pour principale originalité la mise en parallèle de diverses couches convolutives, chacune ayant sa propre paramétrisation (différentes tailles de masque). Les caractéristiques extraites par chaque couche sont ensuite fusionnées via un opérateur dédié.

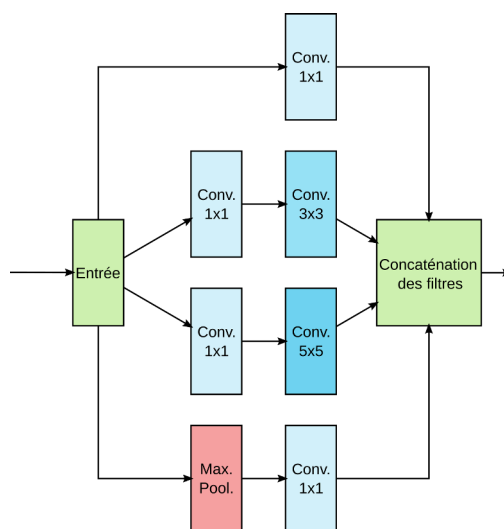


FIGURE 3.3 Illustration de l'*Inception block* proposé par Szegedy et al. pour l'architecture GoogleNet [1]

L'utilisation de ces voies parallèles oriente la recherche vers des réseaux de plus en plus profonds, c'est-à-dire constitués de séquences de couches toujours plus longues. Or, deux problèmes sont rapidement identifiés :

- La composition de fonctions entraîne une évanescence (ou au contraire, une explosion) des gradients sur certaines couches lors de l'opération de rétropropagation. Dans une certaine mesure, cet effet peut être contrôlé par une initialisation judicieuse des poids

- du modèle [43, 44] ou l'utilisation de couches de normalisations intermédiaires [45–47].
- Il a été observé qu'un réseau plus profond est sujet à une dégradation de ses performances [48, 49], non seulement sur l'ensemble de test, mais également sur celui d'apprentissage. Un tel effet est contre-intuitif : ajouter des poids ne devrait théoriquement pas dégrader les performances sur l'ensemble d'apprentissage. En effet, dans le pire des cas, le réseau pourrait se passer des couches supplémentaires en les mapant vers la fonction identité. Cette dégradation n'est donc pas la conséquence d'un sur-apprentissage lié à un excès de paramètres.

Ces observations ont motivé la recherche vers de nouvelles formes d'architectures telles les *Highway Network* [49] ou les *Residual Network* [2]. Toutes deux ont pour objectif d'accroître la profondeur du réseau (jusqu'à plus de cent couches convolutives) tout en résolvant les problèmes liés aux gradients par des biais inductifs facilitant la transmission du signal. Nous détaillons ici la dernière ; ses performances l'ayant rendu très populaire, la rendant aujourd'hui encore très utilisée.

Residual Network He et al. introduisent la notion de connexions résiduelles pour augmenter significativement le nombre de couches du réseau [2]. Le raisonnement des auteurs est le suivant : si un ensemble de couches non linéaires successives peut approximer n'importe quelle fonction $\mathcal{H}(x)$, il lui est tout autant possible d'approximer la fonction résiduelle $\mathcal{F}(x) = \mathcal{H}(x) - x$. La fonction originale se retrouve donc simplement par addition $\mathcal{F}(x) + x$, équation qui est au cœur du *ResNet* : l'entrée d'une (ou plusieurs) couche est additionnée à sa sortie d'où l'appellation de connexion résiduelle. Cette « astuce » permet d'accroître la taille du réseau car cette connexion facilite la propagation des gradients. Par ailleurs, toute section $\mathcal{F}(x)$ du réseau potentiellement inutile pourra converger vers la fonction nulle ; la connexion résiduelle assurant que l'identité sera conservée. Les auteurs démontrent que ces connexions améliorent grandement les performances d'un réseau profond (jusqu'à 101 couches), tout en maintenant celle d'un réseau plus superficiel.

DenseNet Huang et al. étendent le concept des connexions résiduelles à travers une nouvelle architecture nommée DenseNet [50]. À l'inverse du ResNet, pour lequel les connexions résiduelles sont arbitrairement choisies, le DenseNet s'appuie sur un principe de connexion complète : chaque sortie de couche est envoyée en entrée de toutes les couches suivantes appartenant à un même bloc, où chaque bloc est délimité par une opération de pooling. Autrement dit, toutes les couches au sein d'un même bloc possède la même résolution spatiale. Au sein d'un bloc de L couches, il existera donc $\frac{L \cdot (L+1)}{2}$ connexions au lieu des $L + 1$ du ResNet. Autre différence avec ce dernier, la connexion résiduelle ne se fait pas par une addition, mais par

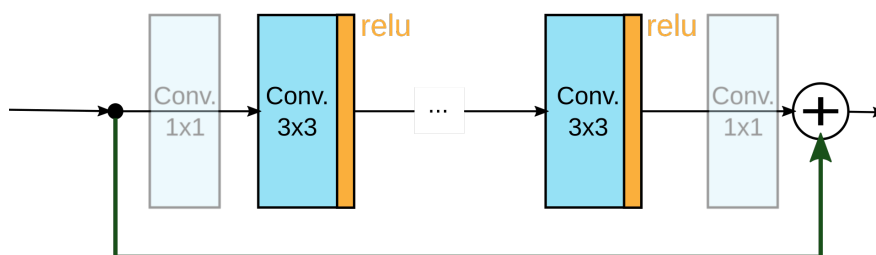


FIGURE 3.4 Connexion résiduelle proposée par He et al. pour l'architecture ResNet [2]. Les convolutions 1×1 placées en entrée et sortie sont optionnelles : elles permettent la projection des cartes de caractéristiques vers des espaces aux dimensions concordantes si nécessaire.

concaténation des cartes de caractéristiques. Pour limiter une sur-paramétrisation du modèle, chaque couche est limitée à douze cartes de sortie, ce qui est significativement plus petit que dans les travaux décrits précédemment. Malgré cela, les auteurs obtiennent des performances supérieures à l'état de l'art et montrent que grâce à ces connexions « denses », le problème des gradients évanescents disparaît. Une des hypothèses fournies pour expliquer cette efficacité repose sur l'idée que la structure dense favorise la réutilisation des caractéristiques entre couches successives, ce qui va dans le sens d'un meilleur signal propagé.

SqueezeNet Inspirés par les premiers modèles d'attention (comme ceux de Wang et al. [51]), Hu et al. [52] proposent une architecture appelée SqueezeNet. Son originalité repose sur le module *Squeeze-É-Excite*. L'idée, là encore, est d'élargir le contexte spatial pris en compte dans une couche convolutive : initialement, le noyau de convolution basique n'a un effet que local, les relations contextuelles à plus large échelle sont donc ignorées. Jusqu'ici, cet effet a été corrigé en accumulant séquentiellement les couches convolutives, mais Hu et al. proposent une autre approche. L'idée est de construire, pour chaque couche, un vecteur contextuel encapsulant les relations globales. Celui-ci est obtenu par *Global Mean Pooling* (c'est-à-dire la moyenne spatiale du tenseur des caractéristiques en sortie de couches), avant d'être fourni à un simple réseau PMC à deux couches qui maintient les dimensions du vecteur. La première étape correspond à la compression des caractéristiques (*Squeeze*), la seconde à leur excitation. Le vecteur résultant est ensuite passé à travers une fonction sigmoïde σ afin de borner ses valeurs dans l'intervalle $[0, 1]$. L'étape d'excitation consiste à multiplier ce vecteur avec le tenseur d'entrée, venant ainsi pondérer chacune des caractéristiques de celui-ci. Cette pondération est souvent associée à la notion d'attention, car elle permet de concentrer le réseau sur certaines cartes de caractéristiques précises.

Ces dernières années, la notion d'attention est devenue centrale dans la conception d'architectures de réseaux de neurones, bien qu'elle ne soit pas toujours construite de la même

manière. À l’instar du SqueezeNet, l’attention peut globalement se comprendre comme un mécanisme de pondération de certaines des caractéristiques de l’entrée de la couche suivie d’une recombinaison de celles-ci. Il est vite apparu qu’un tel concept se prête particulièrement aux applications basées sur des séquences monodimensionnelles (c’est-à-dire le plus souvent des séries temporelles) et notamment celles autour du langage naturel. Parmi les travaux les plus marquants de ces dernières années, Vaswani et al. [53] ont radicalement bouleversé les pratiques du traitement des séquences en proposant un modèle d’attention dit « par Produit Scalaire Normalisé » (*Scaled Dot-Product*). Ce mécanisme a également hérité ultérieurement de l’appellatif « d’Auto-Attention » ; il est au cœur de l’architecture de Vaswani et al. dite *Transformer*. Étant donné le succès obtenu par cette architecture (initialement pour des tâches de traduction), il n’est guère étonnant qu’elle ait été adaptée à la vision par ordinateur ; il faudra néanmoins de nombreuses itérations pour obtenir des performances comparables aux CNNs. Dosovitskiy et al. [6] y parviennent finalement avec une architecture nommée *Vision Transformer* (ViT).

3.1.2 Le modèle du Vision Transformer

Le ViT consiste en une succession de couches d’encodages construites selon la même structure. Chaque couche $\mathbf{E}^n, n \in \{1, \dots, N\}$ (N étant le nombre de couches) reçoit en entrée une séquence $S^n \in \mathbb{R}^{l \times d}$ composée de l éléments (appelés *jetons*) $s_i^n \in \mathbb{R}^d$ où d représente la dimension de chaque jeton. À la différence des architectures de CNNs typiques, où des opérations d’échantillonnage (*pooling*) restreignent progressivement la résolution spatiale de l’entrée, pour le ViT, la taille l de la séquence reste inchangée de couche en couche. Pour la première couche, la séquence est obtenue par une projection linéaire de morceaux d’images (*patches*) vers un vecteur de dimension d . Afin de maintenir l’information de positionnement spatial dans l’image de chaque jeton, on leur additionne un encodage positionnel. Enfin, la séquence ainsi constituée se voit greffer un jeton supplémentaire, dit de *classification*. Telle que représentée sur la figure 3.5, la propagation au sein d’une couche d’encodage se fait en plusieurs étapes :

1. Chaque séquence est d’abord normalisée, utilisant la technique de Ba et al. [47] dite de « Normalisation par Couche » (*Layer Normalization (LN)*).
2. Elle est recomposée via une opération auto-attentive. Celle-ci est réalisée parallèlement K fois, d’où l’appellation originale de *Multi-head Attention (MHA)*.
3. La sortie du MHA est fournie à un réseau de neurones à deux couches, utilisant la fonction d’activation de type *Gaussian Error Linear Unit (GELU)*.

Par ailleurs, le ViT utilise deux connexions résiduelles au sein d'une couche. L'ensemble de ces étapes se résume donc par le jeu d'équations suivant :

$$S_a^n = \text{MHA}(\text{LN}(S^{n-1})) \quad (3.7)$$

$$S_b^n = S_a^n + S^{n-1} \quad (3.8)$$

$$S^n = \text{MLP}(\text{LN}(S_b^n)) + S_b^n \quad (3.9)$$

Il reste à décrire le principe de fonctionnement de la couche MHA. Chaque jeton s_i^n est projeté vers trois composantes distinctes appelées Clé, Requête et Valeur (respectivement *Key*, *Query*, *Value* dans leur nom original) et notées k_i^n , q_i^n , v_i^n . Ces trois vecteurs s'obtiennent par la projection linéaire de chaque jeton d'entrée, ce qui donne, sous forme d'équation matricielle :

$$K^n = I^n \cdot W_k^n \quad (3.10)$$

$$Q^n = I^n \cdot W_q^n \quad (3.11)$$

$$V^n = I^n \cdot W_v^n \quad (3.12)$$

$$(3.13)$$

où $I^n = \text{LN}(S^{n-1})$ représente la séquence de jetons normalisée. Une tête d'attention A^n est obtenue par le produit scalaire normalisé de la matrice de Clé et de Requête :

$$A^n = \frac{Q^n \cdot K^{nT}}{\sqrt{d}} \quad (3.14)$$

Notons que le facteur de normalisation ($\frac{1}{\sqrt{d}}$) qui fait intervenir la dimension de chaque jeton se justifie par des considérations sur la variance du produit scalaire de deux vecteurs aléatoires que l'on souhaite maintenir constante de couche en couche. Depuis A^n , on déduit une matrice de scores d'importance par jeton en la normalisant avec la fonction softmax appliquée ligne-à-ligne, permettant ainsi de pondérer la dépendance entre chaque paire de jetons dans la séquence. Le MHA s'obtient alors comme :

$$S_a^n = \text{softmax}(A^n)V^n \quad (3.15)$$

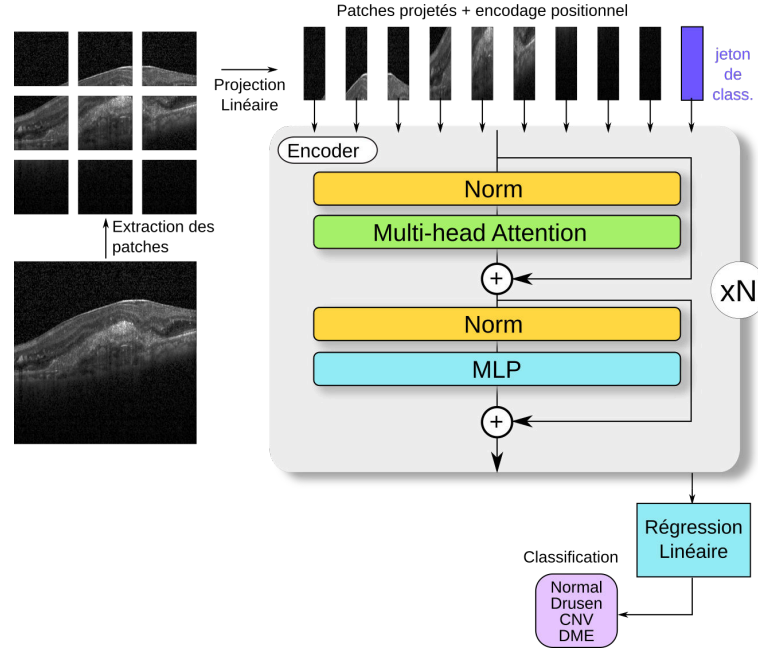


FIGURE 3.5 Représentation d’une architecture ViT utilisée sur la classification d’une image OCT. Le modèle se compose d’une succession de N couches d’encodage. Le jeton de classification de la dernière couche est fourni en entrée d’une régression linéaire dont la sortie représente la décision du modèle.

3.1.3 Variantes du modèle

Le ViT de Dosovitskiy et al. parvient à obtenir la première place au concours de classification *ImageNet* mais requiert pour cela d’énormes quantités de données d’apprentissage, incluant le recours à des bases externes. Une des raisons invoquée pour justifier cette dépendance est l’absence de biais inductifs propres à la vision au sein du modèle, c’est-à-dire de structure inhérente à l’architecture qui favorise une convergence optimale lors de la manipulation d’images. À l’inverse, les CNNs sont naturellement construits pour cette tâche (ils offrent notamment l’invariance en translation et préservent la localité lors de l’extraction de caractéristiques) ; pour cette raison, ils seraient capables de converger vers un meilleur optimum avec moins de données.

Touvron et al. trouve une manière de s’affranchir en partie de cette contrainte sur la taille des bases d’apprentissage : ils démontrent la possibilité d’entraîner le modèle sans utiliser de données supplémentaires (par rapport aux standards d’entraînement des CNNs). Pour cela, ils suggèrent de renforcer l’apprentissage supervisé classique par l’ajout d’un jeton dit de *distillation* dans la séquence. Ce jeton n’est pas entraîné à prédire le label des données mais à la place à reproduire la prédiction faite par un autre modèle (en l’occurrence, un CNN déjà entraîné). Cette forme d’apprentissage distillé est au cœur du modèle DeiT (*Data-Efficient*

image Transformers), qui par ailleurs ne se distingue du ViT que par l’ajout de ce jeton de distillation à la séquence fournie en entrée.

Toujours dans l’idée de favoriser la convergence vers un optimum global, Yuan et al. [54] proposent des modifications plus substantielles à travers leur architecture, le *Tokens-to-Token ViT* (T2T-ViT). Celle-ci introduit la notion de fusion locale de jetons (dans une opération qui rappelle le pooling), permettant de restructurer la séquence de jetons de couche en couche. Cet opérateur se faisant sur les jetons adjacents, il crée une relation de voisinage entre ceux-ci, ce qui résulte bien en un biais inductif en faveur du traitement d’images. Par ailleurs, les auteurs expérimentent sur cette architecture les différentes techniques qui ont contribué aux succès des CNNs (connexions denses, module *Squeeze-Excite*). Une démarche similaire est adoptée par Chen et al. [55] qui proposent le VisFormer, une adaptation du Transformer visant à lui conférer les mêmes biais inductifs qu’un CNN traditionnel. Ces travaux ne représentent qu’une infime fraction de l’activité autour des Transformers (incluant les ViTs), dont l’émergence a initié une transition assez radicale des pratiques en s’éloignant du CNN conventionnel. De nombreuses revues de littérature, y compris consacrées au champ médical [56], ont fleuri sur ce sujet encore nouveau, bien qu’un nombre non négligeable d’entre elles soient encore en révision (par exemple celle très exhaustive d’Azad et al. [57]).

3.1.4 Modèles de segmentation sémantique

Les modèles présentés dans la section précédente ont pour point commun de se limiter à la classification. Autrement dit, pour une image donnée, la sortie de ces réseaux aura la forme d’un vecteur indicateur de la classe prédite. Mais pour de nombreuses applications et tout particulièrement dans le domaine médical, il est souvent nécessaire d’obtenir une granularité plus fine et obtenir une classification à l’échelle du pixel plutôt que de l’image. Cette opération est appelée **segmentation sémantique**. Si les architectures présentées jusqu’ici ne sont pas initialement conçues pour cela, il est aisé de les adapter. Une façon de le faire est d’entraîner le modèle à prédire la classe d’appartenance d’un *patch* (morceau) d’image ; en recombinaison chaque patch, on reconstruit une image segmentée. Conceptuellement similaire, on retrouve cette approche dans les travaux précurseurs de Farabet et al. [58]. L’inconvénient de cette approche est de limiter la prise en compte du contexte à la taille du patch considéré. Farabet et al. résolvent ce problème par l’utilisation des Champs Aléatoires Conditionnels (*Conditional Random Fields*) (CRF), méthode qui suit la théorie statistique des Champs de Markov et qui regroupe les pixels présentant une apparence similaire à travers l’image conditionnellement à une pré-segmentation. Il s’agit donc d’une étape de post-traitement qui a pour avantage de prendre en compte le contexte global de l’image. Le second inconvénient du modèle de Farabet et al. est de produire une segmentation dont la résolution est tributaire de la taille

du patch choisi : plus celui-ci est grand, plus la résolution sera faible. Mais inversement, pour des petits patches, segmenter toute l'image prendra un temps considérablement plus long. Ces deux inconvénients ont poussé au développement d'un nouveau type d'architecture de réseaux de neurones que nous présentons ici.

Réseaux entièrement convolutifs

Long et al. [59] seront les premiers à se passer d'une couche de classification dans leurs travaux sur la segmentation. Plutôt que de reprendre la structure conventionnelle : $N \times$ (Couches convolutives $\rightarrow$ Pooling) $\rightarrow K \times$ Couches totalement connectées $\rightarrow$ Classification (qui décrit plus ou moins toutes les architectures CNNs de la section précédente), les auteurs suggèrent de ne garder que la première étape (les N couches convolutives), afin de conserver la structure spatiale des cartes de caractéristiques que le pooling détruit. Chacun des pixels appartenant à ces cartes est ensuite projeté vers sa classe prédite. Ce faisant, au lieu de passer d'une fonction de coût par image, on obtient une fonction de coût par pixel. La spécificité de cette architecture est donc de n'utiliser que des couches convolutives et de ré-échantillonnage (pooling, unpooling ou interpolation). En revanche, dans les travaux de Long et al. la prédiction finale reste raffinée par l'utilisation des CRF. L'opération de ré-échantillonnage permet d'interpoler les cartes de caractéristiques vers la résolution initiale de l'image d'entrée. Cette opération peut se faire aussi bien par simple interpolation bilinéaire que par déconvolution.

DeepLab-V3 L'architecture DeepLab-V3, introduite en 2017 par Chen et al. [3] franchit un palier technologique, en obtenant de hautes performances malgré l'absence d'un post-traitement de type CRF. Comme son nom l'indique, il s'agit de la troisième itération d'une architecture dont chaque composant a été soigneusement testé. Elle se caractérise par l'utilisation de blocs basés sur l'idée des connexions résiduelles de type ResNet (paragraphe 3.1.1) et surtout sur l'introduction du module ASPP (pour *Atrous Spatial Pyramid Pooling*). La notion de convolutions atrous (à trous) est fondamentale pour cette architecture, en cela qu'elle permet, à moindre frais, la prise en compte d'un plus grand contexte spatial et sans ajouter de paramètres au modèle. Son mécanisme est illustré sur la figure 3.6.

U-Net Introduite par Ronneberger et al. [60], l'architecture U-Net est certainement une des plus emblématiques en ce qui concerne la segmentation sémantique d'images médicales et la plus citée. Outre sa simplicité, un de ses plus gros atouts est sa capacité à générer d'excellentes performances même en ne disposant que de peu de données d'entraînement.

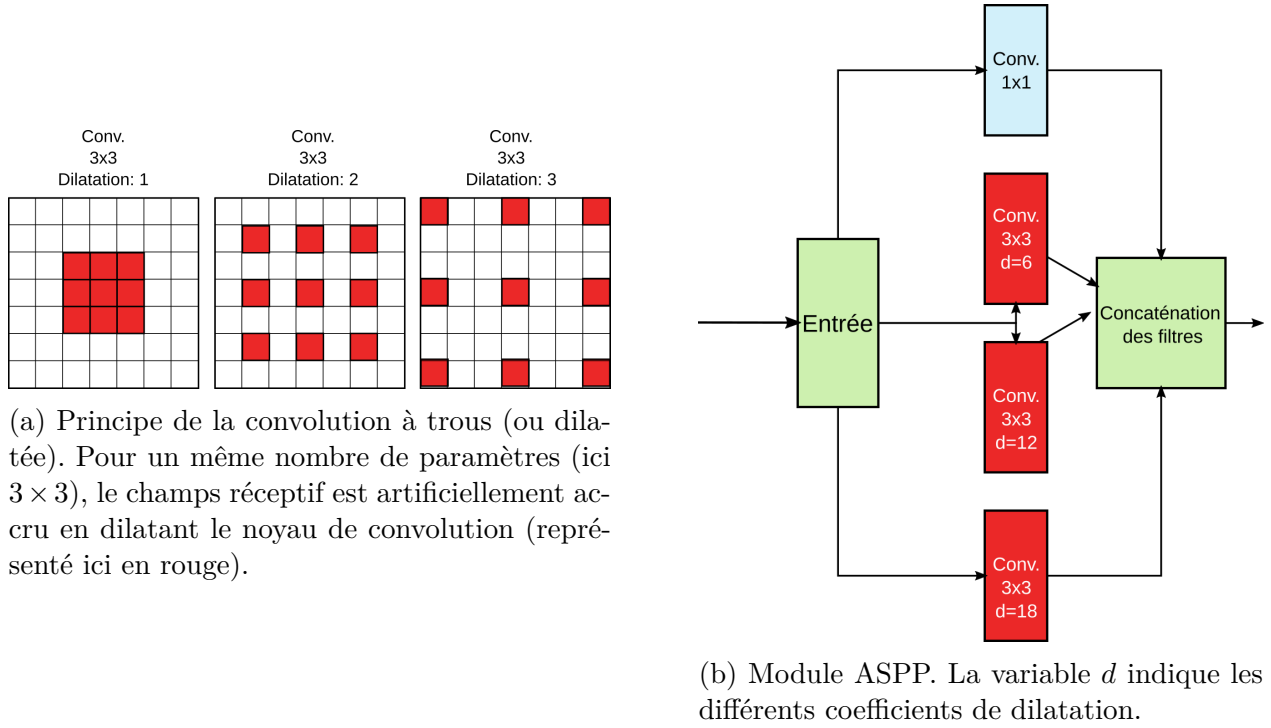


FIGURE 3.6 Illustrations des modules utilisées dans l'architecture DeepLabV3 [3]

Une représentation de l'architecture est fournie sur la figure 3.7. Sa structure respecte le principe, souvent repris depuis, d'encodeur/décodeur. L'encodeur fonctionne similairement à la plupart des architectures présentées jusqu'ici, c'est-à-dire en compressant progressivement les dimensions spatiales de l'image tout en enrichissant le vecteur descriptif à chaque pixel de nouvelles composantes. Le décodeur pour sa part fait l'inverse et ré-interpole progressivement les cartes de caractéristiques jusqu'à la dimension de l'image. Cependant, le U-Net se distingue de ce simple principe en ajoutant des connexions passerelles entre chaque étape de l'encodeur et du décodeur, consistant à concaténer les caractéristiques de l'encodeur à celle du décodeur pour chacune des couches. Celles-ci ont deux objectifs :

- Maintenir intacte la propagation des gradients malgré le nombre de couches du réseau, similairement à ce que fait le ResNet.
- Affiner les détails de segmentation des petites structures. En effet, du fait du principe de sous-échantillonnage/ré-interpolation, les réseaux de segmentation ont tendance à effacer les plus petites structures. En ajoutant ces connexions passerelles, le U-Net corrige cet effet.

Variantes du U-Net L'architecture du U-Net est fondatrice pour sa structure en « U » basée sur un couple encodeur/décodeur symétrique. Peut-être en raison de sa facilité d'im-

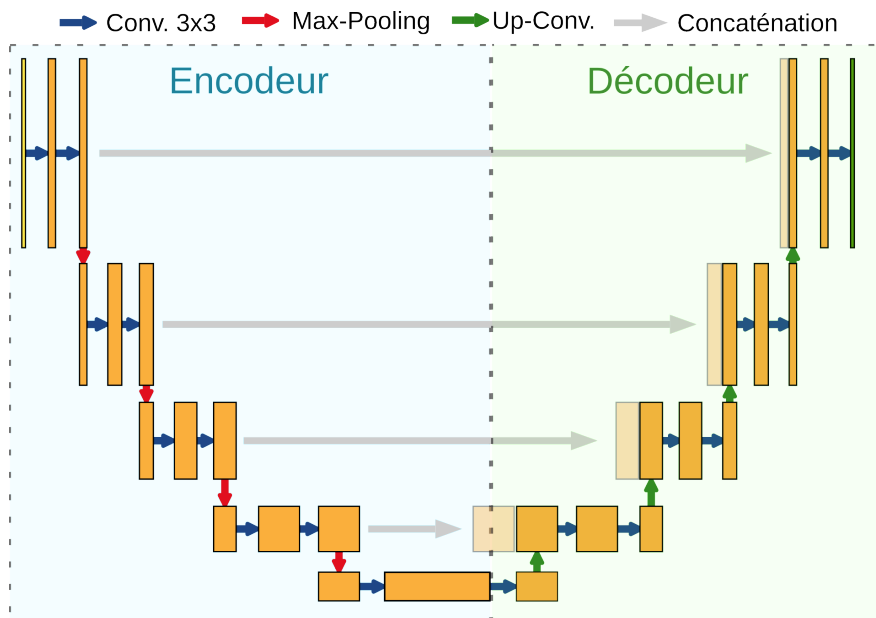


FIGURE 3.7 Architecture originale du U-Net

plémentation, ce concept a inspiré une pléthore de variantes. Bien souvent, les améliorations proposées s'appuient sur des modules empruntés à d'autres architectures pré-existantes. Par exemple, Rundo et al. [61] créent le SE-U-Net en incorporant le module Squeeze-&-Excite dans les différentes couches d'encodeur et décodeur pour la segmentation d'IRM, Oktay et al. [62] ajoutent un mécanisme d'attention sur une variante du U-Net adaptée aux images tridimensionnelles dédiée à la segmentation du pancréas, Zhou et al. [63] étendent le concept des connexions denses (de type *DenseNet*) entre les couches d'encodage et décodage de leur réseau appelé U-Net++.

Il ne s'agit là que de quelques exemples parmi de nombreux autres, qui témoignent parfaitement de la modularité des architectures de réseaux de neurones.

3.1.5 Optimisation et entraînement de modèles

Malgré la diversité des architectures, il est remarquable que celles-ci adoptent essentiellement la même stratégie pour ajuster leurs paramètres lors de la phase d'entraînement. Cette section a pour objectif de clarifier ce mécanisme de manière formelle pour s'accorder sur une définition de l'apprentissage. Précisons que nous nous limitons ici au problème de l'apprentissage supervisé (autrement dit, pour lequel il existe une référence explicite à atteindre, appelée vérité terrain). Bien souvent le cas de l'apprentissage non-supervisé aborde l'entraînement de manière similaire, avec cependant une reformulation de l'objectif visé.

Dans un cas comme dans l'autre, l'entraînement utilise un jeu de données, en pratique souvent de taille conséquente, pour mettre à jour les paramètres du modèle de manière itérative. Ce processus s'arrête une fois la convergence atteinte vers un modèle jugé satisfaisant. Nous le détaillons ici.

Soit un jeu de données $\mathcal{S} = \{(x_m, y_m) \in \mathbb{R}^d \times \{1, \dots, K\}^{d'}\}, |S| = M$, c'est à dire M échantillons (x_m de dimension d), chacun venant avec une classe associée représentée par le label y_m prenant valeur parmi les K classes possibles. Notons que pour des images, en général, $x_m \in \mathbb{R}^{H \times W}$, représentant les deux dimensions spatiales, mais nous nous limitons ici à un espace monodimensionnel sans perte de généralité. La dimension d' du vecteur de label dépend de l'application considérée : ainsi pour la classification d'images, en général $d' = 1$ (une seule classe par image), pour la segmentation sémantique, $d' = d = H \times W$ où H, W représentent la hauteur et la largeur de l'image (une classe par pixel de l'image).

Le réseau de neurones que l'on souhaite entraîner est représenté sous forme d'une fonction $f_\theta : \mathbb{R}^d \rightarrow \{1, \dots, K\}^{d'}$, où θ englobe tous les paramètres (ou poids) du modèle. Cette configuration représente le cas typique rencontré pour l'apprentissage supervisé d'un modèle multiclasse (y compris pour de la segmentation). Un modèle f_{θ^*} idéalement entraîné sur les données $\mathcal{S}$ saura faire une association parfaite entre x_m et sa classe associée par la vérité terrain², c'est-à-dire :

$$f_{\theta^*}(x_m) = y_m, \forall 1 \leq m \leq M \quad (3.16)$$

En revanche, un modèle non-idéal commettra des erreurs. On les quantifie par une fonction de coût (ou d'erreur) notée $\mathcal{L}(f_\theta(x_m), y_m)$ évaluée pour chaque point de $\mathcal{S}$. En la calculant

2. Il s'agit d'une simplification car la sortie d'un réseau de neurones est en réalité continue et non discrète. En réalité, on modélise plutôt la sortie réseau comme une estimation de la probabilité d'appartenance à chaque classe.

sur l'ensemble des données, on obtient une fonction de coût global :

$$J(\theta) = \sum_{m=1}^M \mathcal{L}(f_{\theta}(x_m), y_m) \quad (3.17)$$

Seule la dépendance à θ est indiquée dans la fonction d'erreur J , car il s'agit de la seule variable d'intérêt à optimiser. À titre d'exemple, la fonction d'erreur au carré moyenne (connue sous le nom de *Mean Square Error* (MSE)) mesure directement l'écart entre la sortie du modèle et le label associé à son entrée³ :

$$J(\theta) = \frac{1}{M} \sum_{m=1}^M (f_{\theta}(x_m) - y_m)^2 \quad (3.18)$$

En construisant ainsi une fonction de coût, on peut définir la notion d'entraînement de manière formelle : il s'agit de trouver les paramètres θ^* qui minimiseront $J(\theta)$, c'est à dire qui diminueront l'erreur globale sur l'ensemble des données à disposition :

$$\theta^* = \arg \min_{\theta} J(\theta) \quad (3.19)$$

Le problème de cette minimisation se ramène à un calcul d'optimisation numérique que nous détaillons au prochain paragraphe. Mais ouvrons une courte parenthèse pour préciser qu'il s'agit d'une vision tronquée de l'apprentissage machine, qui en réalité s'appuie sur la théorie de l'apprentissage statistique et justifie l'entraînement d'un réseau comme une tentative de modélisation de la distribution de probabilités décrivant $\mathcal{S}$. Cette interprétation probabiliste sera détaillée à la fin de cette section ; nous fermons pour l'instant cette parenthèse pour revenir à la notion d'optimisation.

Descente de gradients

Abstraction faite de la théorie justifiant cette approche, l'entraînement d'un réseau de neurones se résume à un problème d'optimisation numérique d'une fonction d'erreur. Deux types de solutions existent pour un tel problème :

1. Une solution unique θ^* telle que $J(\theta^*) \leq J(\theta), \forall \theta$ est qualifiée de minimum global.
2. Si la solution identifiée n'est valable que dans un voisinage $\mathcal{N}$ de θ^* , c'est à dire $J(\theta^*) \leq J(\theta), \forall \theta \in \mathcal{N}$, on parle alors de minimum local.

Notons que si la fonction à optimiser J est convexe, les minimum local et global se confondent. Une grande partie de la recherche en optimisation se concentre donc sur les problèmes

3. Le carré garantit la positivité de la fonction

convexes ; mais il est bien connu que les réseaux de neurones n'appartiennent pas à cette catégorie. Idéalement, l'entraînement devrait aboutir à une solution θ^* globale : elle indiquerait ainsi la meilleure configuration de paramètres du modèle parmi l'ensemble des possibilités. En pratique, il est quasiment impossible d'identifier un minimum global et on se contente d'en rechercher un local (il peut y en avoir naturellement plus d'un). Il est aisé de reconnaître un minimum local en faisant quelques hypothèses faibles : il est nécessaire que la fonction de coût global, $J(\theta) : \mathbb{R}^N \rightarrow \mathbb{R}$ soit différentiable presque partout (où N est le nombre de paramètres du modèle). Tout minimum, local ou global, vérifie alors :

$$\nabla J(\theta^*) = 0 \quad (3.20)$$

La démonstration de ce résultat est une application quasi-directe du théorème de Taylor. L'équation 3.20 permet de définir un critère d'arrêt de l'entraînement, correspondant à la découverte d'un minimum local. Partant d'une initialisation θ_0 du modèle, il reste alors à trouver le chemin menant à θ^* (ou du moins une valeur approchée). Pour y arriver, tous les algorithmes d'optimisation génèrent une séquence $\{\theta_k\}_{k=0}^\infty$ et définissent des règles d'évolution. Il existe deux grandes stratégies pour trouver ces itérations. La première a trait à la recherche de ligne directrice (direction à prendre dans l'évaluation des θ_k) tandis que la seconde transforme le problème en la recherche d'une région de confiance (dans l'espace des paramètres) ; toutes deux ont fait l'objet d'importantes recherches ces cinquante dernières années largement documentées par Nocedal J. et Wright S. [64].

L'entraînement d'un réseau de neurones appartient à la première catégorie. À chaque itération k , la fonction de coût est évaluée au point θ_k . On choisit ensuite une direction $p_k \in \mathbb{R}^N$ qui permet de décroître la fonction de coût et on se déplace le long de celle-ci sur une certaine longueur $\alpha_k \in \mathbb{R}^+$ avec pour objectif de remplir ces deux conditions :

$$\begin{cases} \theta_{k+1} &= \theta_k + \alpha_k p_k \\ J(\theta_{k+1}) &\leq J(\theta_k) \end{cases} \quad (3.21)$$

Dans le contexte de l'entraînement d'un réseau de neurones, le scalaire α_k porte le nom de *pas d'apprentissage*.

Toujours en s'appuyant sur la décomposition de Taylor de la fonction de coût, on peut montrer que la direction de plus forte diminution suit le gradient au point θ_k . Dans le cadre de l'entraînement, on choisit donc généralement :

$$p_k = -\nabla J(\theta_k) \quad (3.22)$$

d'où l'appellation « descente de gradients » tire son nom. Notons qu'il ne s'agit pas forcément de la direction optimale pour atteindre le minimum ; uniquement la direction de plus forte descente au point considéré. La méthode de Newton suggère par exemple de calculer la direction :

$$p_k = -\nabla^2 J_k^{-1} \nabla J_k \quad (3.23)$$

Qui fait intervenir la Hessienne inversée de la fonction J au point θ_k . Malheureusement, ni le calcul de dérivées d'ordre supérieur à 1 (pour la Hessienne), ni le calcul de son inverse ne sont en pratique réalisables pour un réseau de neurones. En effet, rappelons que J a pour support $\mathbb{R}^N$, où N représente le nombre de paramètres du modèle. Celui-ci peut se compter en milliers (voire bien plus pour les plus gros modèles). Cette complexité rend inutilisable une grande partie des techniques développées en optimisation numérique dès lors qu'il s'agit d'une architecture un tant soit peu conséquente. De la même manière, il existe une abondante littérature sur le choix optimal du pas d'apprentissage en fonction de la direction choisie [65], mais en pratique, l'entraînement d'un réseau de neurones utilise des heuristiques bien plus simples.

Descente de gradients stochastique

En choisissant $p_k = -\nabla J(\theta_k)$, l'algorithme de descente de gradients peut se réécrire :

$$\theta_{k+1} = \theta_k - \alpha_k \nabla J(\theta_k) \quad (3.24)$$

$$= \theta_k - \alpha_k \nabla \sum_{m=1}^M \mathcal{L}(f_{\theta_k}(x_m), y_m) \quad (3.25)$$

Sous cette forme, il apparaît clairement que le calcul de l'estimateur $J(\theta_k)$ fait intervenir l'ensemble des données pour le calcul des contributions individuelles $\mathcal{L}$, et ce pour chaque itération k d'apprentissage. Dans la pratique, où les données $\mathcal{S}$ peuvent contenir plusieurs milliers ou millions de points, il est bien souvent très inefficace (voire impossible) de calculer le gradient global. On peut en revanche en calculer une estimation. Pour cela, on définit la famille de parties $\{\mathcal{B}_i \subset \mathcal{S} \mid \bigcup_i \mathcal{B}_i = \mathcal{S}\}$ de telle sorte que chaque élément de $\mathcal{B}_i$ est tiré sans remise aléatoirement de $\mathcal{S}$. L'algorithme se base ainsi sur une estimation du gradient global sur un sous-ensemble de données :

$$\theta_{k+1} = \theta_k - \alpha_k \nabla \sum_{m \in \mathcal{B}_i} \mathcal{L}(f_{\theta_k}(x_m), y_m) \quad (3.26)$$

À chaque itération k , un nouveau sous-ensemble $\mathcal{B}_i$ ⁴ est choisi, correspondant donc à un échantillonnage aléatoire des données d'entraînement. On parle donc de descente de gradients stochastiques. Dépendamment d'hypothèses plus ou moins fortes, il est possible de démontrer la convergence uniforme d'un tel processus itératif. Mais nous nous contenterons ici de cette description sommaire, suffisante pour introduire une technique qui appartient en réalité au vaste domaine des algorithmes stochastiques et des martingales, pour lequel il existe de nombreux résultats théoriques fondamentaux [66]. Néanmoins, la version simple présentée ici est bien celle employée pour l'apprentissage des réseaux de neurones.

Propagation inverse des gradients

Jusqu'ici, nous avons présenté la manière de minimiser une fonction de coût à partir de la mise à jour de ses variables par descente de gradients (stochastique ou non). Mais nous avons passé sous couvert la manière dont sont calculés ces gradients pour chacun des paramètres du modèle. Cela implique en particulier de rappeler la dépendance de la fonction de coût à ces paramètres. Remémorons-nous qu'un réseau de neurones $f_\theta : \mathbb{R}^d \rightarrow \{0, 1, \dots, K\}^{d'}$ est une suite d'opérations appliquées à un vecteur d'entrée $\mathbf{x}$ pour produire une sortie y . Chacune de ces opérations implique son propre jeu de paramètres (d'où l'appellation de fonction paramétrique). Par exemple, une couche totalement connectée a pour paramètres un poids $\mathbf{W} \in \mathbb{R}^{\text{in} \times \text{out}}$ et un biais $\mathbf{B} \in \mathbb{R}^{\text{out}}$. Un réseau composé de N couches totalement connectées g_n peut donc s'écrire (on omet le biais par simplicité) :

$$f(\mathbf{x}) = g^N(g^{N-1}(\dots g^1(\mathbf{x}, \mathbf{W}_1), \mathbf{W}_{N-1}), \mathbf{W}_N) = g_{\mathbf{W}_N}^N \circ g_{\mathbf{W}_{N-1}}^{N-1} \circ \dots \circ g_{\mathbf{W}_1}^1(\mathbf{x}) \quad (3.27)$$

Cette représentation du réseau en tant que composition de fonctions est essentielle pour comprendre l'entraînement par rétro-propagation des gradients qui est au cœur de la méthode de mise à jour des poids. Donnons-nous comme outils pour cela le théorème de la dérivation des fonctions composées. Soient f, g deux fonctions à valeur dans $\mathbb{R}$ dérivables sur leur ensemble de définition respectif. Alors on a :

$$(g \circ f)' = (g' \circ f) \times f' \quad (3.28)$$

$$\Leftrightarrow \frac{dg \circ f}{dx} = \frac{dg}{df} \frac{df}{dx} \quad (3.29)$$

4. Appelé *mini-batch* en anglais

Ce théorème se généralise au-delà des fonctions scalaires. Soient $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, $g : \mathbb{R}^m \mapsto \mathbb{R}$ et $h : \mathbf{X} \mapsto (g \circ f)(\mathbf{x})$. Alors on a :

$$\frac{\partial h}{\partial x_i} = \sum_j \frac{\partial h}{\partial g_j} \frac{\partial f_j}{\partial x_i} \quad (3.30)$$

$$\Leftrightarrow \nabla_{\mathbf{x}} h = \left(\frac{\partial f}{\partial \mathbf{x}} \right)^T \nabla_g h \quad (3.31)$$

$\frac{\partial f}{\partial \mathbf{x}}$ est la matrice Jacobienne de f de taille $n \times m$. Or, en représentant un réseau de neurones en composition de fonctions (ses couches), l'application de ce théorème permet de calculer le gradient de chacun de ses paramètres. Qui plus est, une récursivité apparaît naturellement : en effet, l'équation 3.30 sépare le gradient en deux composantes multipliées entre elles :

1. La Jacobienne $\frac{\partial f}{\partial \mathbf{x}}$, qui correspond intuitivement au calcul du gradient de l'opération réalisée à la couche considérée par rapport à son entrée.
2. Le gradient $\nabla_g h$ de la couche suivante.

En poursuivant le raisonnement, le gradient $\nabla_g h$ s'obtient à son tour par le produit de la Jacobienne de g et du gradient de la fonction qui le suit, etc... Autrement dit, pour le calcul du gradient de la couche n , on utilisera le gradient calculé à la couche $n + 1$. L'apparition d'un tel schéma récursif permet de grandement optimiser l'opération du calcul de gradients d'un réseau de neurones. En pratique, pour obtenir cette *rétropropagation* des gradients, l'ensemble des opérations qu'il réalise est traduite sous forme d'un graphe dirigé, où chaque variable issue d'une opération est représentée par un nœud associé à cette opération. Les arêtes pointant vers un nœud représentent les variables utilisées par l'opération du nœud. Cette représentation, outre qu'elle permet un grand nombre d'optimisations (éliminer les nœuds inutiles, ceux qui ne sont pas pris en compte dans le calcul du gradient, etc.) établit un ordonnancement des opérations, depuis le vecteur d'entrée (par exemple une image), jusqu'à la valeur de sortie (en général, le coût associé). $\mathcal{L}$ représente donc le dernier nœud du graphe de compositions du réseau. En notant que naturellement, $\frac{\partial \mathcal{L}}{\partial \mathcal{L}} = 1$, on peut calculer le gradient de tous ses ancêtres en utilisant le théorème de la dérivation des fonctions composées, puis récursivement, des ancêtres des ancêtres, etc. Le grand intérêt de cette approche est la ré-utilisabilité des gradients de la couche $n + 1$ pour calculer ceux de la couche n .

Interprétation statistique

Nous l'avons mentionné plus tôt lors de l'introduction de la notion de fonction de coût, l'apprentissage d'un réseau de neurones (et en réalité, l'essentiel des algorithmes d'apprentissage machine) peut s'interpréter sous le prisme du formalisme probabiliste, qui lui vient à la fois

de sa capacité à établir une mesure (au sens mathématique) sur un ensemble et à l'inhérente stochasticité des données réelles utilisées à l'apprentissage. Nous couvrons dans cette section quelques éléments sommaires établissant les bases de ce formalisme.

On fait l'hypothèse que pour un jeu de données de n éléments $\mathbb{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$, il existe une distribution unique (non accessible et inconnue) $p_{\text{données}}(\mathbf{x})$ d'où ont été tirés les échantillons de manière indépendante et identiquement distribués. Le modèle, lui, prédit une probabilité par échantillons $f(\mathbf{x}, \theta)$. L'entraînement consiste alors à trouver le point $\theta^* \in \Theta$ permettant d'estimer au mieux la vraie distribution des données (on parle d'estimateur statistique).

Pour cela, on introduit la notion de vraisemblance du modèle, définie par :

$$L_n(\theta) = \prod_{i=1}^n f(\mathbf{x}^{(i)}, \theta) \quad (3.32)$$

Elle correspond à la probabilité d'observation jointe de l'ensemble des variables $\mathbf{x}^{(i)}$ sous l'hypothèse qu'elles soient bien indépendantes et identiquement distribuées (I.I.D). On souhaite trouver un modèle capable de maximiser cette probabilité étant données les variables dont nous disposons (qui sont donc bien observées). L'articulation avec la notion de fonction de coût en découle directement : l'apprentissage est la recherche de l'estimateur du maximum de vraisemblance, c'est-à-dire :

$$\theta_n^* = \arg \max_{\theta \in \Theta} L_n(\theta) \quad (3.33)$$

Cette équation se ramène à la minimisation d'une fonction en prenant simplement l'opposé du maximum de vraisemblance. Notons qu'en pratique, plutôt que de manipuler le produit des probabilités, on préfère le substituer par une somme de logarithmes pour des raisons de stabilité numérique :

$$\arg \max_{\theta \in \Theta} L_n(\theta) = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log(f(\mathbf{x}^{(i)}, \theta)) = \arg \min_{\theta \in \Theta} - \sum_{i=1}^n \log(f(\mathbf{x}^{(i)}, \theta)) \quad (3.34)$$

On peut facilement étendre l'équation au cas de l'apprentissage supervisé, pour lequel chaque échantillon $\mathbf{x}$ est accompagné d'un label représentant sa classe. Auquel cas, on souhaite que le modèle représente la distribution de probabilité dans l'espace des classes disponibles conditionnée à l'entrée observée. Autrement dit,

$$f(\mathbf{x}^{(i)}, \theta) = p(\mathbf{y}|\mathbf{x}^{(i)}) \quad (3.35)$$

On définit alors, comme dans l'équation 3.32, la vraisemblance-conditionnelle et on ramène le problème de l'apprentissage à sa maximisation :

$$\theta^* = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log P(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}; \theta) \quad (3.36)$$

3.2 Éléments de la théorie de l'apprentissage statistique

Tout au long des expériences qui seront présentées dans les prochains chapitres, la notion de généralisation a occupé une place prédominante dans les questionnements de recherche. Elle a, selon les auteurs, une définition plus ou moins précise mais peut se résumer de la manière suivante : il s'agit de la capacité d'un modèle à maintenir de bonnes performances sur des données qu'il n'a jamais vu auparavant. Cette notion sera particulièrement discutée dans le chapitre 4, sous un angle empirique et appliqué aux données médicales qui nous intéressent. Mais il s'agit d'une notion fondamentale lorsqu'on parle de modélisation par apprentissage machine et elle occupe de fait une place aux premières loges des recherches théoriques sur l'apprentissage statistique. La difficulté à justifier les performances pratiques des réseaux de neurones a stimulé la curiosité des mathématiciens qui s'y intéressent et en fait un champs de recherche actif faisant appel à des concepts probabilistes avancés. Cette section n'a pas pour ambition de rentrer dans ces détails, mais d'en présenter le contexte dans un cadre définissant les notions de généralisation, de complexité, de capacité...

3.2.1 Généralisation, sous-apprentissage et sur-apprentissage

Pour reprendre l'élégante formulation de Goodfellow et al. [67], le concept de généralisation d'un modèle est ce qui distingue l'apprentissage d'un « simple » problème d'optimisation numérique. La généralisation traduit la capacité d'un modèle à maintenir ses propriétés sur des données qui n'ont pas servi à l'entraînement. Elle se mesure donc sur un ensemble disjoint de données, appelé ensemble d'évaluation ou de test et cette évaluation se fait en pratique à la fin de l'entraînement. Lors de la phase d'entraînement, en minimisant sa fonction de coût, le modèle réduit son nombre d'erreur sur l'ensemble d'apprentissage. Quand elle s'achève, il s'en suit alors une phase d'inférence ; on évalue alors l'erreur du modèle sur les données de test. Trois scénarios peuvent avoir lieu au cours de ces étapes :

- En cas de **sous-apprentissage**, le modèle n'est pas parvenu à réduire son erreur sur l'ensemble d'apprentissage.
- Dans le cas d'un **sur-apprentissage**, le modèle réduit bien l'erreur sur l'ensemble

d'apprentissage mais ne parvient pas à réduire celle sur celui de test.

- Le modèle parvient à la fois à optimiser la fonction de coût, tout en maintenant de hautes performances sur un ensemble de test. On dit alors qu'il **généralise** bien.

Qu'est-ce qui conditionne un modèle à bien, sous ou sur-apprendre ? Cette question est en réalité extrêmement complexe et a fait l'objet de nombreux travaux de recherche, souvent bien antérieurs au succès des réseaux de neurones. Nous définissons ici certains des concepts nécessaires à la compréhension de cette problématique et nous reprenons pour cela les notations de Kawaguchi et al. [68].

Soient $x \in \mathcal{X}$ et $y \in \mathcal{Y}$ l'entrée et la sortie attendue d'un modèle f et $\mathcal{L}$ la fonction de coût utilisée pour entraîner ce modèle. On définit l'espérance de l'erreur (aussi parfois appelée risque espéré ou attendu dans ce contexte) par :

$$\mathcal{R}[f] = \mathbb{E}_{x,y \sim \mathbb{P}(X,Y)}[\mathcal{L}(f(x), y)] = \int_y \int_x \mathcal{L}(f(x), y) \mathbb{P}(x, y) dx dy \quad (3.37)$$

où $\mathbb{P}(X, Y)$ représente la vraie loi de distribution des données. Or, comme celle-ci est inconnue, on se contente de minimiser le risque empirique sur un ensemble d'apprentissage S , défini comme dans la section précédente, c'est-à-dire composé de m points supposés suivre cette distribution :

$$\mathcal{R}_S[f] = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f(x^{(i)}), y^{(i)}) \quad (3.38)$$

De manière formelle, l'écart de généralisation $\mathcal{E}$ est alors défini par :

$$\mathcal{E} = \mathcal{R}_S[f] - \mathcal{R}[f] \quad (3.39)$$

Cette équation représente donc la différence entre la distribution modélisée à partir d'un ensemble de points d'apprentissage et la distribution réelle. En pratique, il est impossible de calculer $\mathcal{E}$ car la distribution réelle $\mathbb{P}(X, Y)$ est inconnue. On se contente donc d'estimer un écart de généralisation en calculant le risque empirique sur un ensemble de test disjoint :

$$\hat{\mathcal{E}} = \mathcal{R}_S[f] - \mathcal{R}_{S_{test}}[f] \approx \mathcal{E} \quad (3.40)$$

Sous l'hypothèse que tous les points soient bien uniquement distribués, et sous condition d'avoir suffisamment de points dans notre ensemble de test, cette approximation converge vers le vrai écart de généralisation.

Le champs de l'apprentissage statistique essaye d'identifier une borne supérieure à $\mathcal{E}$, afin de répondre à la question centrale : quelles propriétés d'un modèle expliquent sa capacité de généralisation ?

On peut également élargir cette question en s'interrogeant sur la cardinalité $n = \text{card}(S)$ (c'est-à-dire la taille de la base d'entraînement) suffisante et/ou nécessaire pour garantir l'efficacité de l'apprentissage.

Il existe bien entendu une infinité de bornes supérieures ; mais plus celle-ci est petite, plus elle est pertinente. On peut citer plusieurs résultats sur les propriétés spécifiques aux réseaux de neurones qui permettent d'obtenir une borne supérieure à la généralisation estimée : P.L. Barlett [69] démontre que pour certaines architectures bien spécifiques, la norme L1 des poids du réseau intervient dans l'expression de la borne, Golowich et al. [70] étendent ce résultat à des architectures plus complexes en utilisant différentes normes sur les matrices qui composent chaque couche. Ces travaux nécessitent d'introduire un formalisme qui n'est pas propre aux réseaux de neurones, mais qui intervient dès lors qu'on explore les propriétés des bornes supérieures de généralisation. Nous proposons ici d'en donner un aperçu, à commencer par la notion d'hypothèse « Probablement Approximativement Correcte » (PAC).

Hypothèse PAC Étant donné un ensemble S , il est fréquent de trouver plus d'un modèle qui minimisera le risque empirique (sémantiquement, on préfère le terme d'*hypothèse* dans ce contexte).

$$\mathcal{R}_S[f_1] = \mathcal{R}_S[f_2] = 0 \quad (3.41)$$

Si l'hypothèse $h \in \mathcal{H}$ vérifie $\mathcal{R}_S[h] = 0$, on dit qu'elle est consistante sur S . Tout au mieux, on espère que l'hypothèse sélectionnée (dans l'espace de toutes les hypothèses $\mathcal{H}$) minimise également le risque espéré. Pour cela, on introduit la notion d'approximativement correct : si $\mathcal{R}[h] \leq \epsilon$ (idéalement avec ϵ tendant vers 0), on dit que h est approximativement correct. Enfin, il faut encore noter que les points tirés de S le sont aléatoirement et de manière finie, il existe donc un écart potentiel pouvant provenir de l'estimation approximative de la distribution réelle. L'estimation d'erreur ϵ dépend donc également de la confiance $1 - \delta$ à s'approcher de la distribution réelle à partir des échantillons finis. Auquel cas, on ne peut qu'espérer que la probabilité que le risque espéré du modèle soit inférieur à ϵ soit elle-même supérieure à $1 - \delta$, autrement dit :

$$P(\mathcal{R}[h] \leq \epsilon) \geq 1 - \delta \quad (3.42)$$

où δ est un paramètre de confiance dans l'hypothèse h . Sous ces deux conditions, on dit que h est **Probablement Approximativement Correcte** (PAC). S'il existe un algorithme permettant de trouver une hypothèse $h \in \mathcal{H}$ qui respecte cette équation, on dit que $\mathcal{H}$ est PAC apprenable. La notion de PAC est fondamentale dans la mesure où la plupart des bornes supérieures de généralisation démontrées pour une hypothèse h ou une classe d'hypothèses $\mathcal{H}$ ne sont en

réalité que probablement et approximativement correct (c'est-à-dire que toute borne ϵ dépend d'un degré de confiance δ). Nous pouvons dès lors introduire plusieurs concepts qui permettent de dériver ces bornes à partir de considérations intrinsèques sur une classe d'hypothèses. Une des manières d'approcher le problème est d'identifier le pouvoir séparateur de l'hypothèse en question, c'est-à-dire sa capacité à distinguer deux points distincts, indépendamment de leur position dans l'espace et de la classe qu'on leur assigne. Cette capacité est quantifiée par la dimension de Vapnik-Chervonenkis.

Dimension de Vapnik-Chervonenkis Soit un ensemble de points, associés chacun à une classe (0 ou 1). On dit qu'une famille $\mathcal{H}$ de fonctions disperse cet ensemble, si, quel que soit le label associé à chaque point, il existe une fonction de $\mathcal{H}$ qui puisse correctement classifier chaque point. On appelle dimension de Vapnik-Chervonenkis (ou dimension VC) de la famille $\mathcal{H}$ sur un ensemble D le nombre entier k tel que :

- tout ensemble de k points de D est complètement dispersable par $\mathcal{H}$;
- il existe un jeu de $k + 1$ points non complètement dispersable par $\mathcal{H}$

Par exemple, la dimension VC de $\mathcal{H} = \{f : x \rightarrow ax + b, (a, b) \in \mathbb{R} \times \mathbb{R}\}$ (l'ensemble des fonctions affines) est de 3. Géométriquement, ce résultat s'interprète par le fait que quelque soit le label étiqueté sur 3 points quelconques, il est toujours possible de faire passer une droite qui les sépare de telle sorte à maintenir les points de même classe du même côté de la droite.

Appliquée à un réseau de neurones, la dimension VC quantifie le plus grand nombre de points de la base d'apprentissage que le réseau est capable de classifier correctement, qu'importe la classe de chaque point. En 1989, Blumer et al. [71] relient la notion de classe de concepts (c'est-à-dire l'espace des classes associées aux points X , tel que $C \subseteq 2^X$) et la notion d'hypothèse PAC et démontrent qu'il existe une hypothèse PAC sur C si et seulement si la dimension VC de C (notée $d_{VC}(C)$ et correspondant à la cardinalité de l'ensemble des points $X_n \subseteq X$ dispersé par C) est finie.

Cela fournit une application concrète de la dimension de Vapnik-Chervonenkis utilisée dans le cadre de la théorie sur la généralisation. Étant donnée une famille de modèle $\mathcal{H}$ associée à une classification binaire, et soit la fonction de coût dite 0/1 :

$$R_S[f] = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(f(X_i) \neq Y_i) \quad (3.43)$$

qui est un estimateur de la probabilité d'erreur du modèle f , on peut montrer que la probabilité que l'écart de généralisation soit supérieure à ϵ est bornée par :

$$P(\sup_{f \in \mathcal{H}} |R_S[f] - R(f)| > \epsilon) \leq 8d_{VC}(\mathcal{H}, n)e^{-n\epsilon^2/32} \quad (3.44)$$

où $n = \text{card}(S)$ et $d_{VC}(\mathcal{H}, n)$ est le nombre de points dispersables par tout modèle $f \in \mathcal{H}$ quelle que soit la classe associée aux n points $x_1, \dots, x_n$. Ce résultat est plus fréquemment fourni sous la forme :

$$\mathbb{E}[\sup_{f \in \mathcal{H}} |R_S[f] - R(f)|] \leq 2\sqrt{\frac{\log d_{VC}(\mathcal{H}, n) + \log 2}{n}} \quad (3.45)$$

L'équation 3.45 correspond bien à une borne supérieure de l'erreur de généralisation commise par une classe de modèles, qui dépend des propriétés intrinsèques de la classe d'hypothèses (à travers d_{VC}) et du nombre de points d'apprentissage.

Pragmatiquement, elle a cependant peu d'applications directes. En effet, l'estimation de la dimension VC de $\mathcal{H}$ est pratiquement impossible au delà de quelques cas d'architectures triviales. En particulier, les travaux de Zhang et al. [72] ont montré qu'une architecture classique de réseau de neurones est capable de mémoriser parfaitement un ensemble d'entraînement conventionnel y compris en randomisant les labels associés à chaque point. A priori, ce résultat témoigne donc d'une très grande dimension VC pour un tel modèle. Auquel cas, il est clair qu'une borne supérieure basée sur celle-ci est trop grande pour obtenir une estimation pertinente de la généralisation.

D'autres bornes potentiellement plus strictes ont également été proposées, reposant sur d'autres mesures sur l'espace des hypothèses. En particulier, mentionnons la complexité de Rademacher (empirique ou non), qui fournit un cadre théorique justifiant les travaux de Zhang et al [72]. Sa définition est la suivante :

Soit un ensemble de points $S = (x_1, x_2, \dots, x_m)$ suivant une distribution $\mathbb{P}(X)$ et une classe d'hypothèses $\mathcal{H}$, la complexité empirique de Rademacher est définie par :

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma[\sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(x_i)] \quad (3.46)$$

où $\sigma = (\sigma_1, \dots, \sigma_m)$ est un vecteur aléatoire défini de telle sorte que les σ_i prennent valeurs dans $\{-1, 1\}$, sont indépendants et suivent une distribution de Bernoulli ($p = 0.5$). On peut reconnaître dans cette expression la forme d'une fonction de coût associée à une hypothèse f et des labels (aléatoires) σ_i . Cette analogie permet de la justifier de manière plus intuitive : la complexité empirique de Rademacher mesure en moyenne la plus forte corrélation possible

entre un échantillon f de la classe d'hypothèses $\mathcal{H}$ sur un ensemble de points S à qui on associe des classes aléatoires σ_i . Notons que cette définition ne fonctionne que dans le cas de la classification binaire, il en existe cependant des extensions plus générales.

On peut démontrer que, quelle que soit une fonction $f \in \mathcal{H}$ et avec une probabilité d'au moins $1 - \delta$ (c'est-à-dire dans un contexte PAC), on a :

$$\mathbb{E}[f(x)] - \frac{1}{m} \sum_{i=1}^m f(x_i) \leq 2\hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \quad (3.47)$$

En particulier, si f est une fonction de coût, on retrouve une borne théorique sur l'erreur de généralisation sur la classe d'hypothèses $\mathcal{H}$.

Nous n'avons donné ici qu'un petit aperçu des outils théoriques inventés pour cadrer la notion de généralisation. Pour plus de détails, nous nous référons à des ouvrages de référence, dont celui de M. Rostamizadeh [73] qui propose une description approfondie de ces notions et des démonstrations rigoureuses des quelques résultats mentionnés ici.

3.2.2 Distribution multimodale et adaptation de domaines

Dans le cadre de nos travaux sur la segmentation d'images rétinienues, la notion de généralisation au sens strict ne permet pas de cadrer de manière pertinente la discussion sur les performances de nos modèles. En effet, comme le montreront nos résultats, la littérature est suffisamment riche pour qu'il soit désormais possible d'obtenir des modèles de qualité⁵ sur des ensembles de données fixés (pour lesquels on peut faire l'hypothèse que les points de l'ensemble sont indépendants et identiquement distribués). Cette hypothèse IID est bien entendu fondamentale derrière le concept de généralisation, car elle justifie la mesure empirique du risque comme approximation du risque espéré. Or, en réalité, il est courant que cette hypothèse ne soit pas respectée ; plus exactement, si l'indépendance est bien respectée, ce n'est pas forcément de l'identité de la distribution dont sont issus les échantillons. Par exemple, certaines applications reposent sur des données issues de différentes sources d'acquisitions : on parle alors d'apprentissage multimodal ; Ramachandram et Taylor [74] proposent une revue de littérature sur ce sujet spécifique dans le cadre des architectures de réseau de neurones. Cet exemple à lui seul justifie donc la nécessité de repenser la théorie sur la généralisation en adoptant l'hypothèse de plusieurs distributions distinctes modélisant les données. Plus précisément, ce changement de distribution entre bases de données peut avoir lieu à plusieurs niveaux :

5. La notion de qualité peut paraître subjective : dans nos travaux, nous la quantifions par l'utilisation de diverses métriques et nous hiérarchisons les scores obtenus relativement à l'état de l'art.

- Les entrées $\mathcal{X}$ du modèle ne suivent pas toutes la même distribution. Dans le cas médical, on peut par exemple supposer que la distribution dépend du protocole d’acquisition d’images, du groupe ethnique propre à chaque base de données, etc... Pour une même modalité, on obtient de fait des images de distributions très diverses.
- Des sorties $\mathcal{Y}$ constituant la vérité terrain cible du modèle. En pratique, ce cas de figure se rencontre dès lors que les données n’ont pas été annotées de manière homogène : par exemple, les définitions de classes peuvent différer entre bases de données, ou, dans le cadre de la segmentation, le style d’annotations peut varier entre annotateurs.
- Ces deux précédentes conditions peuvent également se combiner, rendant encore plus complexe la modélisation de la distribution via diverses sources de données forcément limitées.

Plusieurs travaux abordent ce sujet sous des angles relativement différents. Nous nous restreignons à ceux ayant pour sujet la segmentation sémantique, car il s’agit de notre application dans le chapitre 4. Dans le cadre du premier scénario (changement de distribution de $\mathcal{X}$), plusieurs travaux suggèrent de faire de l’adaptation de domaine, en distinguant le domaine *source* (en général, celui pour lequel on dispose de nombreuses données) et le domaine *cible* (celui visé lors du déploiement du modèle). Ganin et al. [75] proposent d’apprendre une forme de représentations invariantes au changement de domaines en introduisant des couches de standardisations des caractéristiques suivant une approche adversariale : l’idée est d’entraîner un modèle externe, appelé discriminateur, à distinguer les caractéristiques internes en fonction de leur domaine d’origine. Les poids du modèle de segmentation sont mis à jour dans la direction opposée au discriminateur pour obtenir la représentation invariante. Hoffman et al. [76] ont une approche similaire, mais en adaptant cette fois à la fois l’espace latent des caractéristiques et les images elles-mêmes. Leur méthode utilise également une approche adversariale pour générer les modifications, suivant le modèle du CycleGAN de Zhao et al. [77]. Une approche transverse à ces méthodes repose sur l’apprentissage auto-supervisé. Il s’agit d’utiliser les prédictions, appelées pseudo-labels, d’un premier modèle pour en entraîner un second. Ces prédictions sont réalisées sur les images du domaine cible, sur lesquelles on ne dispose de peu ou pas d’annotations. Cette approche a été notamment expérimentée par Zou et al. [78] et par Zheng et Yang [79]. Le modèle qui génère les pseudos-labels est généralement appelé « maître » et celui entraîné avec « l’élève ». Ce dernier est mis-à-jour suivant la descente de gradients conventionnelle, tandis que le maître est obtenu par la moyenne glissante des poids de l’élève.

On peut noter que ces approches sont surtout pratiques lorsqu’on souhaite standardiser la sortie d’un réseau dans un contexte où il existe des images appartenant à des distributions distinctes $\mathcal{X}^{(i)}$. En quelque sorte, l’action porte donc sur une correction à l’entrée du réseau.

D'autres travaux se concentrent sur une autre problématique qui est celle de l'existence de différents domaines $\mathcal{Y}^{(j)}$. Dans le cadre de la segmentation sémantique, et encore plus dans le contexte médical, l'hypothèse généralement adoptée est qu'il existe une vérité terrain unique y mais que pour une raison ou une autre, les annotations contiennent des perturbations dues à une erreur de l'annotateur ou à une incertitude sur l'interprétation des frontières de structures. Cette annotation « bruitée » est généralement notée $\hat{y}$. Récemment, Luo et al. [80] ont étudié le comportement d'un modèle de segmentation entraîné sur une vérité terrain bruitée et ont montré qu'un réseau, en particulier dans les premières phases de l'apprentissage, tend à détecter la vraie distribution y avant de sur-apprendre sur $\hat{y}$. Il s'agit cependant d'un travail purement théorique sur des données artificielles et relativement triviales à segmenter. Pour une vision exhaustive des travaux pratiques portant sur le sujet, nous recommandons la revue de littérature proposée par Song et al. [81]. Elle fournit un tour complet d'horizons des principales techniques d'apprentissage dans le cadre d'annotations bruitées, organisée suivant quatre catégories : l'implémentation d'architecture intrinsèquement robustes au bruit, l'utilisation de régularisation lors de l'apprentissage, la conception d'une fonction de coût insensible au bruit et la sélection pondérée des échantillons lors de l'entraînement. Voici quelques exemples de ces techniques développées dans des travaux divers : l'approche par « Apprentissage Confiant » [82] illustre le principe d'architectures robustes. Elle vise à identifier l'origine du bruit en estimant explicitement la distribution jointe $Q_{y,\hat{y}}$. Différentes techniques sont comparées pour cela, mais l'intérêt final est de fournir la matrice de transition de bruits, modélisant la probabilité $T_{ij} = p(y = i | \hat{y} = j)$. L'ajout d'une couche d'adaptation aux bruits en fin de réseau permet donc de pondérer la prédiction en fonction de l'estimation de la probabilité d'annotations erronées dépendamment de la classe prédite. Xu et al. [83] se servent de cette technique pour entraîner une architecture de segmentation sémantique sur des données grossièrement annotées.

Li et al [84] introduisent une autre approche reposant sur le principe de régularisation via la fonction de coût : ils proposent une méthodologie permettant de combiner de bonnes annotations avec celles bruitées : un modèle est d'abord entraîné sur les annotations de qualité ; puis utilisé sur les images annotées grossièrement. En utilisant le critère du coût le plus faible (Gui et al. [85]), les auteurs identifient des régions où l'annotation bruitée est la plus probablement correcte. En générant des régions (dites « superpixels »), une approche par graphe est ensuite utilisée pour rassembler celles se ressemblant le plus (similairement à ce que font les CRFs).

3.3 Comprendre un réseau de neurones

Cadrer le concept de généralisation permet d’anticiper le comportement d’un modèle sur un ensemble de données inconnues. Malheureusement, mesurer les performances d’un modèle n’est pas synonyme d’une compréhension de celui-ci : une ou plusieurs métriques permettent de se faire une idée de l’efficacité moyenne d’un modèle mais non de son fonctionnement. Par ailleurs, si le modèle manque de transparence dans sa prise de décision, il existe des applications pour lesquelles des statistiques d’ensemble ne suffisent pas à obtenir la confiance de l’utilisateur. Ces notions de transparence et de confiance peuvent sembler aussi bien abstraites que subjectives. Elles ont cependant des implications concrètes : dans le domaine médical par exemple, elles sont régulièrement invoquées pour justifier la difficulté de déploiement des réseaux de neurones. En effet, il est important de savoir si un modèle est biaisé envers certaines populations souvent minoritaires dans les données d’entraînements. À défaut de constituer des bases de données pour chacune des populations que l’on souhaite tester -il existe quasiment une infinité de sous-ensembles minoritaires possibles au sein d’une population-, la question du mode de fonctionnement du modèle devient primordiale. Elle agite de plus en plus la communauté scientifique : à l’heure d’un déploiement de plus en plus massif d’algorithmes d’apprentissage machine dans divers pans de la société, les mécanismes mis en œuvre dans leur prise de décision se doivent d’être le plus transparent possible. Les réseaux de neurones qui mettent en jeu des millions de paramètres et produisent des fonctions non-linéaires extrêmement complexes, posent problème à cet égard car intrinsèquement et par leur échelle ils échappent à toute compréhension intuitive.

Proposer des méthodes permettant à l’utilisateur de saisir leur mécanisme de fonctionnement est l’objet du champs de recherche ayant trait à l’explicabilité et l’interprétabilité des modèles d’apprentissage. Ces deux termes (aux contours imprécis, nous y reviendrons) interrogent les chercheurs bien au delà du champs de l’informatique. L’action d’expliquer, dans le but de faire comprendre, de démontrer ou de persuader relève intrinsèquement de la cognition humaine et donc aussi bien du champs de la philosophie, des mathématiques, des neurosciences que de la psychologie. De quelle manière un modèle peut-il expliquer sa prise de décision ? Une compilation d’articles telle que proposée dans [86] permet de mesurer l’importance de ce questionnement sur des champs d’applications aussi variés que l’analyse causale des modèles génératifs ou encore en psychologie dans le cadre de la recherche des biais des algorithmes de recrutement de candidats. Dans le cadre médical qui nous concerne davantage, une compréhension hypothétique d’un modèle présente de nombreux avantages :

- Naturellement, elle aide à anticiper ses biais vis-à-vis d’une population donnée. Plus généralement, un modèle qui explique sa prise de décision est plus aisé à améliorer par

son concepteur car ce dernier peut le déboguer en retraçant les raisons des erreurs.

- Elle renforce l’acceptabilité clinique du modèle en permettant au médecin de réfuter ou d’accepter les arguments qu’il lui fournit en complément de sa prédiction de diagnostic.
- Elle permet potentiellement de découvrir de nouvelles manières de diagnostiquer une maladie.

Un très grand nombre d’articles ont été publiés proposant des approches permettant d’expliquer les sorties d’un modèle. En les recensant, plusieurs travaux ont proposé des taxonomies sous le format de revues de littérature, spécifiques au domaine médical [87–98] ou plus générales [99–103]. Face à cette abondance, T. Speith [104] propose même une méta-analyse des taxonomies existantes. Tous ces travaux ne sont pas redondants mais témoignent de la difficulté à définir des notions relevant in fine de la psychologie humaine. En effet, une explication est un processus menant à la compréhension, mais cette dernière, à mi-chemin entre la perception et le raisonnement, est également une notion aux contours mal définis. Il n’existe pas de consensus sur ce qui rend un modèle compréhensible, ni sur les stratégies à adopter pour y arriver. Malgré ces ambiguïtés, la notion d’interprétabilité est mise très régulièrement en avant comme la grande marche à franchir dans le développement des nouveaux modèles pour faciliter leur déploiement.

3.3.1 Interprétabilité ou explicabilité ?

Cette question semble présupposer d’une complémentarité entre ces néologismes ou du moins qu’ils désignent deux notions bien distinctes. En réalité, il n’existe pas là non plus de définition arrêtée pour ces termes et la plupart des revues citées précédemment proposent la leur. Plusieurs auteurs (Tjoa et Guan [88], Nauta et al. [105]) revendiquent de les utiliser de manière interchangeable. À l’inverse, d’autres comme Lipton Z. [106] porte un regard très critique sur le flou sémantique qui les encadre. En réponse, Barredo Arrieta et al. [107] tentent de les distinguer conceptuellement : l’interprétabilité d’un modèle serait une propriété passive, inhérente à ce dernier. Pour un tel modèle, son fonctionnement est compréhensible naturellement par un observateur humain. Certains modèles seraient ontologiquement compréhensibles : les k-plus proches voisins, la régression ou classification linéaire, l’arbre de décision... Cette propriété leur viendrait du lien direct qui unit l’entrée du modèle (non ou peu transformée par celui-ci) et sa sortie : c’est donc la simplicité de cette connexion qui permet son interprétation humaine.

Inversement, l’explicabilité est une caractéristique active qui « désigne toute action ou procédure suivie par le modèle dans l’intention de clarifier ou détailler ses fonctions internes ». Selon cette définition, un modèle est donc interprétable ou non de manière inhérente, là où l’explicabilité est un processus extérieur au modèle qui permet de le rendre interprétable alors

qu'il ne l'est pas initialement. Précisons que d'autres définitions pour ces notions existent (par exemple dans les travaux de Montavon et al. [108]). Comme l'observent Barredo Arrieta et al, la force d'une explication est toujours dépendante de l'audience à laquelle elle s'adresse. Une explication jargonneuse pourra être utile voire convaincante pour un spécialiste armé du savoir nécessaire pour la comprendre mais n'éclairera pas la compréhension du néophyte. Ce constat suggère que dans le cycle de vie d'un modèle, le processus d'explicabilité est pluriel, fonction de l'audience à laquelle on souhaite s'adresser. Là encore, prenons l'exemple fictif du déploiement clinique de nos réseaux de neurones, et illustrons les besoins d'explicabilité à travers quatre cas de figure.

Cas A Les autorités de régulation sanitaire seront certainement intéressées par une compréhension des structures apprises sur l'ensemble des données par le modèle, autrement dit la distribution statistique qu'il modélise et plus généralement son espace de représentations. Elles pourront également exiger que l'anonymat des données soit garanti et qu'il soit donc impossible de retracer l'origine des données d'entraînement pour quiconque aurait accès au réseau. Répondre à une telle exigence fait nécessairement appel à une description du mécanisme interne du modèle.

Cas B Dans l'optique d'un déploiement efficace, le concepteur du modèle sera pour sa part intéressé par l'optimisation de la taille de son modèle et voudra donc pouvoir estimer l'importance de chacun des paramètres, ou bien le pouvoir de discrimination linéaire de chacune des couches de caractéristiques. L'explicabilité est alors perçue comme une dissection du modèle permettant de comprendre chacune de ses sous-parties. Sous ce prisme, la construction de la prédiction d'un modèle peut-être analysée comme une forme de phénomène émergent, un modèle n'étant que la somme de ses parties, il suffit d'expliquer celles-ci une-à-une.

Cas C Plutôt que ces considérations générales, l'utilisateur en clinique (patient ou praticien) sera probablement plus intéressé par une explication de ce que le modèle a regardé dans l'image qui le concerne pour prendre sa décision. Cet aspect de l'interprétabilité -qui ne traite pas de la représentation de la globalité des données mais d'une seule décision- appartient au champs de recherche sur la génération d'attribution locale.

Cas D Enfin, le biologiste souhaitera peut-être connaître les liens de causalité qu'identifie un modèle, par exemple en s'interrogeant sur l'importance de tel ou tel biomarqueur pour son diagnostic sur une population donnée. On peut illustrer ce dernier point en citant Arcadu et al. [109] qui utilisent la rétro-propagation guidée des gradients pour identifier les marqueurs

primitifs de progression de la rétinopathie diabétique dans l'image de fond d'œil.

Ce déploiement fictif illustre la pluralité des exigences qui existe en terme d'explicabilité d'un modèle. Cet exemple permet également de contextualiser nos propres travaux, dont la portée relève essentiellement du cas C (et D dans une moindre mesure).

3.3.2 Modèles intrinsèquement interprétables

Dans la vaste gamme des modèles de classification existants, certains sont considérés comme interprétables de façon inhérente. Ce degré d'interprétabilité intrinsèque n'est pas une propriété absolue ni même quantifiée, mais plutôt une sorte d'appréciation collective de la communauté scientifique vis-à-vis de ces modèles. On compte parmi eux les algorithmes d'arbre décisionnel ou les modèles linéaires, même si l'interprétabilité de ces derniers dans un espace de très grandes dimensions est discutable [110]. D'autres approches hybrides intègrent des mécanismes interprétables au sein de modèles complexes qui ne le sont pas a priori. Les cartes d'attentions, utilisées au sein des CNNs ou des Transformers et qui servent à pondérer l'importance de certaines régions de l'image d'entrée, sont également généralement perçues comme appartenant à cette catégorie.

3.3.3 Attributions locales post-hoc

Les méthodes par attributions locales n'ont pas vocation à expliquer un modèle mais davantage sa prédiction associée à une entrée donnée. Dans le cadre de la vision par ordinateur, il s'agit de générer des cartes dites d'attributions indiquant les zones de l'image d'entrée qui ont influencé le modèle dans sa prédiction et c'est généralement l'approche la plus empruntée en ce qui concerne l'explicabilité des modèles médicaux appliqués aux images. En effet, elles se prêtent théoriquement à l'identification des marqueurs locaux que les médecins lisent sur une image pour faire leur diagnostic (tumeur, lésions, occlusions, etc...). Cependant, cet espoir d'un modèle capable de réaliser un diagnostic correct et d'identifier les structures locales causales est sévèrement critiqué par Ghassemi et al. [111], pour qui les méthodes d'explicabilité ne peuvent servir de justification à un diagnostic individuel. Pour en arriver à cette conclusion, ils génèrent les cartes d'attributions issus de plusieurs modèles (entraînés ou non) produisant une prédiction correcte (ou non) : ces cartes se révèlent bien souvent extrêmement similaires, ce qui tend à contredire qu'elles puissent servir de justificatif à la prédiction. Par ailleurs, les cartes produites, outre qu'elles peuvent aussi être erronées, nécessitent elles-même une interprétation de la part du médecin, ce qui tend à les rendre caduques. En revanche, Ghassemi et al. ne contestent par leur utilité : par exemple, elles peuvent permettre d'identifier des biais du modèle sur un ensemble de données. Un cas typique est documenté par

Winkler et al. [112] : un modèle entraîné à la détection du cancer de la peau se révèle (après attribution locale) se concentrer davantage sur les traces de chirurgie que sur les lésions cutanées : le modèle se contente donc d'identifier les patients déjà traités plutôt que la maladie en elle-même. Un tel biais majeur peut donc s'éviter par la génération d'attributions locales sur des ensembles variés de populations. Nos travaux proposent une contribution originale à ce champs de recherche, nous détaillons donc ici les principales techniques d'attributions locales post-hoc.

Occlusion

En 2014, Zeiler et al. [113] introduisent la manière la plus simple d'expliquer un réseau de neurones : masquer une à une ses entrées et observer comment varie sa sortie. Dans le cadre de réseaux convolutifs, l'approche par occlusion consiste à utiliser un masque rectangulaire pour cacher, par fenêtre glissante, les différentes parties de l'image. Les régions cachées de l'image qui modifient le plus la sortie du réseau sont susceptibles d'être celles sur lesquelles il se base pour faire sa prédiction.

Gradients et rétropropagation guidée des gradients

L'une des toutes premières approches pour générer une attribution locale tire parti de la différentiabilité d'un réseau de neurones. L'idée est simplement de calculer la fonction de coût associée à une prédiction et de la dériver par rapport au point d'entrée. Pour une image, cette dérivée correspond à une carte de même dimension, appelée « saillance » (*Saliency*) par ses concepteurs (Simonyan et al. [114]). Intuitivement, elle indique les pixels dont la variation influe le plus l'erreur du modèle.

Springenberg et al. [115] élaborent à partir de cette idée de base et introduisent la rétropropagation guidée, qui consiste en une simple modification du principe de la saillance. Pour chacune des fonctions non-linéaires empruntées lors de la rétropropagation, seules les valeurs de gradients les plus élevées sont conservées. Cependant, ces deux méthodes sont connues pour produire des résultats bruités.

Cartes d'Activation par Classe

À la différence des deux précédentes, la méthode CAM (pour *Class Activation Mapping*) proposée par Zhou et al. [116] ne repose pas sur un calcul de gradients. Elle s'applique sur un type d'architecture bien spécifique : celles utilisant une pénultième couche convolutive suivie d'une fonction agrégatrice de type *Global Mean Pooling* (GMP) et d'une couche de

classification linéaire. Son fonctionnement est simple et relativement intuitif :

Soit $\mathbf{x} \in \mathbb{R}^{N \times H \times W}$ le tenseur créé par cette couche convolutive et $\mathbf{W} \in \mathbb{R}^{N \times C}$ les poids de la dernière couche (celle de classification). La sortie $\mathbf{y} \in \mathbb{R}^N$ produite par l'agrégateur est obtenue par la prise de la moyenne du tenseur le long des dimensions spatiales :

$$y^{(n)} = \frac{1}{H \times W} \mathbf{x}_{h,w}^{(n)} \quad (3.48)$$

La vecteur de prédiction $\mathbf{p} \in \mathbb{R}$ s'obtient alors par $\mathbf{p} = \mathbf{W}^T \mathbf{y}$. En injectant l'équation 3.48 dans cette dernière, on peut réécrire la probabilité associée à une classe c particulière sous la forme :

$$p_c = \sum_{n=1}^N w_c^{(n)} y^{(n)} = \sum_{n=1}^N \sum_h \sum_w w_c^{(n)} \mathbf{x}_{h,w}^{(n)} = \frac{1}{H \times W} \sum_h \sum_w \sum_{n=1}^N w_c^{(n)} \mathbf{x}_{h,w}^{(n)} \quad (3.49)$$

Cette équation est obtenue simplement en réordonnant les sommations des équations précédentes. Elle met en évidence la contribution de chacune des cartes appartenant à $\mathbf{x}$ sur le score de la classe ciblée p_c , pondérée par les poids de la couche linéaire.

L'algorithme CAM ne fait donc que calculer $\sum_{n=1}^N w_c^{(n)} \mathbf{x}_{h,w}^{(n)}$, c'est-à-dire la somme pondérée des caractéristiques spatiales issues de la dernière couche convolutive. La carte résultante a pour dimension $H \times W$ et préserve la disposition spatiale des éléments que le réseau a observé pour tirer sa prédiction. En ce sens, il s'agit bien d'une attribution locale. En revanche, elle a pour inconvénient d'être généralement de très faible résolution. En effet, la dernière couche convolutive arrive en général après plusieurs étapes de pooling ayant drastiquement réduit la résolution spatiale des caractéristiques.

Grad-CAM

L'inconvénient de la méthode CAM est de reposer sur un type d'architecture bien précis. Si notre modèle n'utilise pas de fonction d'aggrégation spatiale finale ou bien la fait suivre de plusieurs couches connectées au lieu d'une, le principe sous-jacent aux CAM ne peut s'appliquer. Selvaraju et al. [117] proposent une solution simple pour répondre à ce problème. À l'instar de CAM, l'idée est de trouver un vecteur de pondération des différentes cartes de caractéristiques issues d'une couche convolutive (pas nécessairement la dernière). Une fois ce vecteur trouvé, la carte d'attribution s'obtient par la somme pondérée des caractéristiques de la couche et s'interprète de la même façon que CAM. Comme précédemment, notons $\mathbf{x}$ la sortie de la couche convolutive choisie. La carte d'attribution est obtenue par :

$$\text{Attr} = \sum_{n=1}^N \alpha_c^{(n)} \mathbf{x}^{(n)} \quad (3.50)$$

Il reste à trouver la pondération $\alpha_c^{(n)}$ pertinente. Selvaraju et al. proposent, similairement à ce qui se fait pour la saillance, de calculer la dérivée de la prédiction du réseau par rapport aux caractéristiques $\mathbf{x}$ et d'aggréger spatialement celle-ci. On obtient alors :

$$\alpha_c^{(n)} = \frac{1}{H \times W} \sum_h \sum_w \frac{\partial p_c}{\partial \mathbf{x}_{h,w}^{(n)}} \quad (3.51)$$

Pour chaque carte de $\mathbf{x}$, la dérivée permet d'estimer sa contribution à la prédiction du modèle. De part sa simplicité, l'approche Grad-CAM est l'une des méthodes de génération d'attributions locales les plus populaires. Elle corrige certains des inconvénients de CAM (notamment sa dépendance à un type d'architecture) en tirant parti des avantages de la rétropropagation. Cependant, elle a aussi ses limites : si la méthode Grad-CAM est calculée sur une des couches profondes, elle souffre du même travers que CAM, à savoir qu'elle génère une carte d'attribution de faible résolution. Mais à l'inverse, le choix d'une couche plus superficielle ne permet pas d'extirper des structures complexes et limite donc la dimension interprétable de la méthode. La nécessité d'un compromis apparaît donc. En pratique, ce sont quasiment exclusivement les couches les plus profondes qui sont utilisées.

LRP et Décomposition de Taylor Profonde

En 2015, Bach et al. [118] mettent en place un nouveau formalisme pour générer des cartes d'attributions locales, non spécifique aux réseaux de neurones. En effet, il s'applique à tout modèle pouvant se décomposer en succession d'opérations se représentant sous forme de nœuds d'un graphe de calcul communiquant entre eux par message ; description qui de facto sied particulièrement bien aux architectures neuronales. La transmission entre nœuds d'un certain score indique la contribution de chacun à la sortie du modèle. Pour chaque nœud, ce score s'obtient comme la somme pondérée du score de ses ancêtres. L'algorithme consiste alors à faire remonter ce score depuis la sortie du modèle jusqu'à son vecteur d'entrée. La pertinence (c'est-à-dire le score en question) s'interprète alors comme la contribution individuelle à la prédiction de chacun des éléments de ce vecteur. L'algorithme de Bach et al. porte le nom de Layer-wise Relevance Propagation (LRP) (Rétropropagation de Pertinence par Couche). Pour avoir un sens interprétable, la pertinence se voit doter d'un certain nombre de propriétés. Soient $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$ la fonction qui représente le modèle étudié et $x \in \mathbb{R}^d$ une image, composée de p pixels, R la carte de Pertinence que l'on souhaite obtenir (et R_p la valeur de cette carte au pixel p).

- La Pertinence doit être conservative, c'est-à-dire que $\sum_p R_p(x) = f(x), \forall x$.
- Elle doit être positive en tout point : $\forall x, p, R_p(x) \geq 0$.

L'algorithme LRP, plutôt que chercher le lien direct entre la valeur de Pertinence en entrée et celle en sortie, va la décomposer progressivement en la propageant de couche en couche, à partir de la sortie du modèle et en remontant jusqu'à son entrée. À chaque couche, le score propagé maintient ses propriétés (conservation et positivité). Ainsi, si on note d^l le nombre de neurones pour la couche (l) d'un réseau, la propriété de conservation stipule :

$$f(x) = \dots = \sum_{i=1}^{d^{(l+1)}} R_i^{(l+1)} = \sum_{i=1}^{d^{(l)}} R_i^{(l)} = \dots = \sum_{i=1}^p R_i^{(1)} \quad (3.52)$$

Bien que ces conditions restreignent le choix de $R^{(l)}$, il n'en reste pas moins qu'une infinité de fonctions les respectent. Bach et al. ajoutent d'autres conditions relativement intuitives pour restreindre l'espace de recherche : par exemple, les activations les plus fortes qui unissent deux neurones (lors de la phase d'inférence, c'est-à-dire la phase de propagation normale) doivent correspondre aux contributions les plus fortes en terme de rétro-propagation de la pertinence. Plusieurs heuristiques sont donc suggérées en fonction du type de modèle.

Mais plutôt que de les détailler, nous préférons nous référer aux travaux plus récents de Montavon et al. [119, 120] qui unifient l'approche LRP et celles basées sur les gradients (saillance et variantes) au sein d'une même théorie reposant sur la Décomposition de Taylor Profonde (*Deep Taylor Decomposition*). Pour cela, rappelons que la décomposition de Taylor d'une fonction multivariée f au voisinage d'un point $\tilde{\mathbf{x}}$ est donnée par :

$$f(\mathbf{x}) = f(\tilde{\mathbf{x}}) + \left(\frac{\partial f}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\tilde{\mathbf{x}}} \right)^\top \cdot (\mathbf{x} - \tilde{\mathbf{x}}) + \epsilon \quad (3.53)$$

L'idée de Montavon et al. consiste à supposer qu'il existe un point $\tilde{\mathbf{x}}$ racine de f pour lequel :

$$\begin{aligned} f(\mathbf{x}) &= 0 + \left(\frac{\partial f}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\tilde{\mathbf{x}}} \right)^\top \cdot (\mathbf{x} - \tilde{\mathbf{x}}) + \epsilon \\ &= \sum_p \frac{\partial f}{\partial x_p} \Big|_{\mathbf{x}=\tilde{\mathbf{x}}} \cdot (x_p - \tilde{x}_p) + \epsilon \\ &\equiv R_p(\mathbf{x}) + \epsilon \end{aligned}$$

Si ϵ est petit, la fonction R_p est approximativement conservative. Elle est par ailleurs positive par construction au voisinage de $\tilde{\mathbf{x}}$ et respecte ainsi les deux propriétés attendues du score de pertinence. Malheureusement, identifier une racine de f proche de $\mathbf{x}$ n'est pas trivial. En revanche, la décomposition de Taylor démontre la possibilité de reconstituer une fonction localement comme la somme de contributions de ces variables d'entrées. Ce constat inspire Montavon et al. à utiliser la Décomposition de Taylor Profonde couche par couche redéfinis-

sant formellement les règles de l'algorithme LRP. Supposons que la fonction de pertinence existe, pour chaque neurone x_j de la couche $l+1$. On peut alors l'exprimer sous forme d'une somme de contributions qu'apportent les neurones x_i de la couche l connectés à x_j :

$$\begin{aligned} R_j &= \sum_i \frac{\partial R_j}{x_i} \Big|_{\tilde{x}_i} \cdot (x_i - \tilde{x}_i) + \epsilon_j \\ &= \sum_i R_{ij} \end{aligned}$$

Or, dans le sens de la propagation avant, (donc de la couche l vers $l+1$), l'algorithme LRP établit déjà que la fonction de pertinence du neurone x_i doit respecter :

$$R_i = \sum_j R_{ij} \quad (3.54)$$

Mises ensemble, ces deux équations aboutissent à une expression explicite de la pertinence du neurone x_i de la couche l en fonction de celle de la couche $l+1$ qui la précède :

$$R_i = \sum_j \frac{\partial R_j}{x_i} \Big|_{\tilde{x}_i} \cdot (x_i - \tilde{x}_i) + \epsilon_j \quad (3.55)$$

Il reste cependant à identifier la racine $\tilde{x}_i$, qui dépend elle de la forme de R_j . Dans le cas d'une couche totalement connectée classique avec activation de type ReLU (c'est-à-dire $x_j = \max(0, \sum_i x_i w_{ij} + b_j)$), Montavon et al. identifient formellement la racine la plus proche de x_i et démontrent que la fonction de pertinence s'obtient alors comme :

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_k w_{kj}^2} R_j \quad (3.56)$$

Cependant, cette solution s'obtient dans l'hypothèse que la racine $\tilde{x}_i$ n'est soumise à aucune contrainte sur son support de définition. En pratique, ce n'est souvent pas le cas et les valeurs de x_i sont bornées sur un intervalle. Par exemple, si le neurone x_i est obtenu en sortie de ReLU, son support est $\mathbb{R}^+$; s'il s'agit du vecteur d'entrée, ce sera plutôt $[0, 255]$ (pour une image) ou $[0, 1]$. La forme de R_i dépend donc du support du neurone x_i . Montavon et al. distinguent différentes règles de définition de R_i en fonction de ces contraintes. Le calcul des LRP tel que présenté est donc plus complexe que les différentes techniques vues jusqu'ici (il est nécessaire d'implémenter différentes règles de rétropropagation en fonction de chaque opérateur), mais cette technique obtient généralement de bien meilleures performances que celles présentées précédemment. Notons qu'elle ne repose pas sur une rétropropagation automatique des gradients, mais plutôt sur une estimation locale de la dérivée du score de

pertinence.

Dans le cas des Transformers, deux opérations nécessitent de redéfinir les règles de propagation : l'addition et la multiplication de vecteurs. La première est utilisée pour assurer les connexions résiduelles tandis que la seconde sert à calculer la matrice d'attention. Chefer et al. [121] dérivent les équations permettant de propager le score de Pertinence pour ces architectures tout en respectant les propriétés de conservation.

Gradients Intégrés

Sundararajan et al. [122] proposent une alternative simple aux LRPs appelée Gradients Intégrés. Cette méthode fonctionne sur tout type de réseaux sans avoir à fournir des ré-implémentations spécifiques à chaque opérateur. Les Gradients intégrés (IG) nécessitent tout d'abord de définir une référence x' (pour les applications sur des images, les auteurs préconisent un simple fond noir). À partir d'une image x , l'algorithme IG mesure la contribution cumulée (calculée en tant que saillance) de chacun des éléments compris sur le chemin qui relie x à x' . Ainsi, en notant x_i le i -ème élément de x , IG calcule :

$$\text{IG}(x)_i = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \cdot (x - x'))}{\partial x_i} d\alpha \quad (3.57)$$

La spécificité des travaux de Sundararajan et al. est de définir un ensemble de quatre axiomes définissant les attributs souhaitables d'une méthode de génération d'attributions. Ils peuvent être résumés ainsi :

1. Invariance à l'implémentation : Si deux modèles ont exactement le même comportement, alors ils doivent générer la même attribution quelle que soit l'entrée.
2. Complétude : La somme des attributions sur la variable d'entrée est égale à la différence entre celle-ci et le point de référence.
3. Sensibilité : Si la sortie d'un modèle ne dépend pas d'une des composantes de son entrée, l'attribution qui lui est associée doit être nulle.
4. Linéarité : L'attribution obtenue à partir de la combinaison linéaire de deux modèles doit être la même combinaison de leur attribution individuelle.

Bien entendu, les auteurs montrent que seule l'approche par Gradients Intégrés permet de satisfaire ces quatre axiomes.

3.4 Métriques d'évaluation des performances

Dans le cycle de vie de la conception d'un modèle, l'évaluation de celui-ci, une fois entraîné, est de toute première importance. En particulier, dans le domaine médical où un déploiement clinique est potentiellement visé, comprendre et analyser les forces et faiblesses du système est indispensable. Malheureusement, il existe une tendance à parfois survoler cet aspect, ce qui est d'ailleurs avancé comme une des potentielles raisons du déploiement relativement restreint des réseaux de neurones en clinique (Shah et al. [123]). Dans cette section, nous proposons une présentation synthétique des questionnements autour de l'évaluation des performances d'un modèle, que ce soit de la sélection de la métrique à l'application contextuelle de celle-ci. Pour cela, nous nous appuyons sur le rapport très détaillé de Maier-Hein et al. [124] (dont il existe une version synthétisée [125]) qui fournit une excellente description de l'ensemble des métriques dédiées à la vision par ordinateur et surtout de leurs limites en fonction de l'application considérée. Ce travail (issu d'un consortium international ayant conduit à de nombreux ateliers, sondages et réunions d'experts) est né du constat de l'existence de nombreux problèmes dans la littérature scientifique sur le traitement d'images médicales : ambiguïtés dans le choix d'une métrique voire absence de justification, interprétation biaisée des résultats, erreur d'implémentation... Les auteurs regroupent ces problèmes en trois catégories :

- Une formulation discutable du problème à traiter. Par exemple, plutôt que de la segmentation de lésions, il peut être judicieux de formuler la tâche sous forme d'un problème de détection, dépendamment de ce qui est nécessaire d'un point de vue clinique.
- Une mauvaise sélection de métriques : certaines ne sont tout simplement adaptées pour juger de la qualité d'un modèle. Par exemple, une mesure d'exactitude (*accuracy* en anglais) n'est guère judicieuse dans le cadre d'un déséquilibre très fort de classes.
- Une application erronée de la métrique, dont l'exemple typique est d'ignorer l'origine des données, en agrégeant par exemple des performances sur des images doublons (ou prises sur le même patient par exemple).

Il faut bien noter que, en bien ou en mal, l'usage fait loi en ce qui concerne l'utilisation des métriques. Fut-elle adaptée ou non, si telle métrique est un standard historique sur un problème donné, son utilisation devient incontournable dans la pratique pour permettre une comparaison juste et évolutive des techniques existantes dans l'état de l'art. Rien n'empêche cependant de les discuter et de proposer une métrique alternative en complément.

Pour Maier-Hein et al., les métriques sont regroupées en quatre types, dépendamment de l'application considérée : classification d'images, segmentation sémantique, détection d'objets et segmentation d'instances. Cette dernière application n'est qu'une combinaison des

deux qui la précèdent. C'est également la seule qui n'apparaîtra pas dans nos travaux présentés dans les prochains chapitres et nous l'éluons donc pour cette raison. Bien entendu, conceptuellement, une même métrique peut apparaître dans plusieurs catégories car il existe une porosité naturelle entre elles (la segmentation sémantique n'est qu'une forme de classification, à l'échelle du pixel). Enfin, on distingue le cas binaire (par exemple : malade/sain) et le cas multi-classes (par exemple : sain/stade 1/stade 2/...), mais comme nous le verrons, tout problème multi-classe peut se binariser.

3.4.1 Scores de classification

Deux cas de figure se distinguent : dans le premier, le modèle émet une sortie discrète et dénombrable, correspondant à la classe prédite. Dans le second, le modèle sort une probabilité d'appartenance à une classe (et donc par définition non dénombrable).

Cas fini dénombrable Soit la vérité terrain constituée de la séquence d'entiers $t_1, t_2, \dots, t_N$ et $y_1, y_2, \dots, y_N$ la prédiction correspondante d'un modèle, $t_i, y_i \in \{1, \dots, K\}$, K étant le nombre de classes. Pour les métriques sur un tel ensemble dénombrable, le calcul se fait généralement à partir d'une matrice de confusion $\mathcal{M}$ de taille $K \times K$. On construit celle-ci suivant la règle :

$$\mathcal{M}_{i,j} = \sum_{k=1}^N \mathbb{I}(t_k, i) \mathbb{I}(y_k, j) \quad (3.58)$$

Chaque ligne représente la vérité terrain par classe et chaque colonne la prédiction du modèle par classe. Dans le cas binaire $K = 2$, on définit un ensemble de variables telles que représentées dans le tableau 3.1. Ces définitions de VP, VN, FP, FN, P et N, permettent de

construire les métriques suivantes :

$$\text{Sensibilité} = \frac{VP}{P} \quad (3.59)$$

$$\text{Spécificité} = \frac{VN}{N} \quad (3.60)$$

$$\text{Précision} = \frac{VP}{VP + FP} = \frac{VP}{PP} \quad (3.61)$$

$$\text{Prévalence} = \frac{P}{P + N} \quad (3.62)$$

$$\text{Score F1} = \frac{2VP}{2VP + FP + FN} \quad (3.63)$$

$$\text{Exactitude} = \frac{VP + VN}{VP + VN + FN + FP} = \frac{VP + VN}{N + P} \quad (3.64)$$

$$\text{CCM } \phi = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \quad (3.65)$$

TABLEAU 3.1 Matrice de confusion dans le cas binaire

		Classes prédites	
Total (P+N = PP + PN)		Prédiction Positive (PP)	Prédiction Négative (PN)
Classes	Positif (P)	Vrai positif (VP)	Faux négatif (FN)
	Négatif (N)	Faux positif (FP)	Vrai Négatif (VN)

La compréhension et l'intérêt de ces différentes métriques s'affinent lorsqu'on en étudie le comportement à la limite de leurs différentes composantes, ce qui correspond à la simulation d'un déséquilibre de classe. Certaines notamment se révèlent être des indicatrices peu pertinentes de la performance dans le cas d'une prévalence fortement déséquilibrée, c'est-à-dire quand la classe positive est faiblement représentée. Illustrons-le avec le cas du passage à la limite $P \rightarrow 0$ pour un modèle « bête » qui ne prédirait que la classe négative ($PP = 0$). La spécificité et l'exactitude tendent alors vers 100% tandis que la sensibilité tend vers une forme indéterminée de type 0/0. Dans ce scénario bien précis, des scores comme la précision ou le score F1 sont plus adaptés pour juger de la qualité d'un modèle. Une solution est de reporter non pas un seul score mais un couple de métriques complémentaires, telles la sensibilité et la spécificité (qui permettent donc d'identifier à la fois le taux de vrai positif et le taux de vrai négatif, soit l'ensemble des cases de la matrice de confusion normalisée par ligne).

Définies de la sorte, ces métriques ne s'appliquent qu'au cas binaire. En réalité, il est bien souvent nécessaire de les étendre à un cas multi-classes. Une approche typique (mais non la seule) consiste à utiliser la moyenne « macro » : le problème multi-classes est redéfini en un

ensemble de cas binaires (un par classe). La classe considérée est alors définie comme positive et toutes les autres négatives (voir la figure 3.8 qui illustre la binarisation par classe) ; de là, on se ramène aux calculs des métriques binaires. On obtient ainsi un ensemble de scores (un par classe), dont on prend la moyenne pour obtenir un score unique. Celle-ci peut être pondérée pour prendre en compte le déséquilibre de classes ou simplement pour accentuer les classes les plus importantes.

Il existe d'autres métriques adoptant une approche probabiliste du problème de la mesure. On peut par exemple structurer le problème sous forme d'une estimation de probabilités d'accord entre deux observateurs. C'est le cas du coefficient κ de J. Cohen [126] défini comme :

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (3.66)$$

où P_o est la proportion d'accord observé et P_e la proportion d'accord dans le cas où les deux observateurs feraient des prédictions de manière aléatoire (en respectant la distribution de classes). Plutôt que de reprendre les notations du tableau 3.1, on introduit par simplification la matrice de contingence plus générique suivante (Tableau 3.2) :

TABLEAU 3.2 Matrice de contingence générique entre deux observateurs pour un problème de classification binaire

		Observateur 1	
		Positif	Négatif
Observateur 2	Positif	a	b
	Négatif	c	d

P_o n'est rien d'autre que l'exactitude et s'obtient donc par :

$$P_o = \frac{a + d}{a + b + c + d} \quad (3.67)$$

P_e est un peu plus subtil à calculer car il faut déterminer la probabilité que les deux observateurs se soient entendus tout en ayant fait leur prédiction au hasard. On calcule la probabilité que l'observateur 1 associe aléatoirement la classe « positif » à un échantillon tiré sur les $T = a + b + c + d$ en jeu :

$$P_1^+ = \frac{a + c}{T} \quad (3.68)$$

Similairement pour l'observateur 2, sa probabilité d'associer la classe positif à un échantillon est de :

$$P_2^+ = \frac{a + b}{T} \quad (3.69)$$

En supposant l'indépendance entre les observateurs, la probabilité qu'ils associent donc la même classe « Positif » au hasard est donc :

$$P^+ = P_1^+ \times P_2^+ \quad (3.70)$$

Le raisonnement tient également sur les probabilités marginales pour la classe « Négatif ». On obtient donc de la même manière :

$$P^- = P_1^- \times P_2^- \quad (3.71)$$

La probabilité d'accord aléatoire résulte de l'addition de la probabilité d'accord par classe :

$$P_e = P^+ + P^- \quad (3.72)$$

Cette interprétation de la performance d'un modèle en tant que mesure de l'accord entre observateurs présente un certain nombre de propriétés élégantes : elle corrige naturellement le déséquilibre de classes (en pondérant l'accord d'observation par l'accord obtenu par tirage aléatoire), elle se généralise très facilement au cas multiclasse (il suffit d'étendre le calcul des probabilités marginales pour chaque classe et de les additionner à l'instar de l'équation 3.72) et elle peut également s'interpoler dans le cas d'un plus grand nombre d'observateurs. Pour ce cas précis, A. Conger [127] propose par exemple de mesurer les κ deux-à-deux et d'en prendre une moyenne (on parle de *kappa de Light*). Krippendorff et al. [128] étend son principe de calcul de manière plus formelle pour le cas multi-observateur en reprenant le formalisme des équations du cas à deux observateurs (on parle alors de *alpha de Krippendorff*). Une application de cette métrique est proposée dans le chapitre 5.2.3 avec pour objectif de mesurer l'accord entre plusieurs médecins ayant participé à un sondage consistant à associer un score à plusieurs images.

Cas continu La plupart des réseaux de neurones, du fait d'être nécessairement différentiables, produisent une réponse continue. En utilisant une opération de type softmax ou sigmoïde, cette réponse s'interprète généralement comme la probabilité qu'associe le modèle à chacune des classes. La classe recevant la plus forte probabilité (dans le cas binaire, la classe associée à $p > 0.5$) est celle prédite. Dans une certaine mesure, cette probabilité représente une mesure de confiance du modèle dans sa prédiction. Précisons que cette hypothèse est cependant contestée, par exemple par Gal et al. [129] qui proposent une autre définition de l'incertitude du modèle. Quoiqu'il en soit, un modèle A qui prédit correctement une classe avec $p_A = 0.95$ est intuitivement plus performant qu'un modèle B faisant la même prédiction

		Groundtruth				
		0	1	2	3	4
Predicted	0	TP	FP	FP	FP	FP
	1	FN	TN	TN	TN	TN
	2	FN	TN	TN	TN	TN
	3	FN	TN	TN	TN	TN
	4	FN	TN	TN	TN	TN

(a) Classe 0

		Groundtruth				
		0	1	2	3	4
Predicted	0	TN	TN	TN	FN	TN
	1	TN	TN	TN	FN	TN
	2	TN	TN	TN	FN	TN
	3	FP	FP	FP	TP	FP
	4	TN	TN	TN	FN	TN

(b) Classe 3

FIGURE 3.8 Définition de la matrice de confusion dans le cas multi-classes. Les images sont issues de captures d'écran prises sur notre outil d'interprétation des métriques de classification. Cet outil adopte une convention différente (transposée de la matrice de confusion). Les notions de Vrai Positif (*True Positive (TP)*), Vrai Négatif (*True Negative (TN)*) dépendent de la classe considérée.

correcte avec $p_B = 0.6$. Or les métriques que nous avons décrites jusqu'à présent ne font pas cette distinction (dès lors que les prédictions entre A et B sont les mêmes).

Pour ce faire, nous introduisons deux métriques, reposant sur un postulat similaire. Nous décrivons celles-ci dans le cas binaire, mais comme nous l'avons vu, tout problème multi-classes peut se ramener au cas binaire et les métriques continues ne dérogent pas à cette règle.

Aire sous la courbe ROC (*Area Under the Receiver Operating Characteristic*) (AUROC) La sensibilité et la spécificité permettent de décrire entièrement la matrice de confusion normalisée par ligne d'un modèle. Dans le cas idéal du classifieur parfait, cette matrice de confusion normalisée est simplement l'identité. Inversement, un classifieur qui se trompe systématiquement obtient la matrice miroir de celle-ci. Dans le premier cas, le couple sensibilité/ spécificité vaut (100%, 100%) et (0%, 0%) pour le second. A priori, un modèle correctement entraîné obtiendra un couple de scores se situant quelque part entre ces deux extremums. Cela dépendra du seuil de confiance qu'on souhaite associer à la prédiction. Soit $p^{(+)}$ la probabilité associée à la classe positive émise par le modèle. Dans le cas binaire et par défaut, si $p^{(+)} > 0.5$, on considère que la classe prédite est positive. La valeur de 0.5, bien qu'intuitive (c'est celle qui sépare bien l'espace des probabilités en deux valeurs égales), est arbitraire. La condition de prédiction de la classe positive peut ainsi se formuler sous la forme $p^{(+)} > s$, où s est un seuil arbitraire : plus il est élevé, plus le nombre de prédictions positives diminue (la sensibilité aura tendance à diminuer) et le nombre de prédictions négatives augmente (la spécificité aura tendance à augmenter). Pour obtenir des métriques évoluant dans le même sens, on définit la fonction d'efficacité du récepteur -dite courbe ROC- comme :

$$\text{ROC}(s) = (\text{sensibilité}(s), (1 - \text{spécificité})(s)) \quad (3.73)$$

On peut condenser cette fonction paramétrique en un seul score en calculant l'aire sous la courbe de la fonction, aussi appelée AUROC :

$$\text{AUROC} = \int_{s=0}^1 \text{ROC}(s) ds \quad (3.74)$$

Sur la figure 3.9, nous fournissons quelques exemples de courbes ROC typiques suivant la qualité du modèle. Le graphique 3.9d représente notamment un modèle qui prédit systématiquement la classe opposée à celle attendue.

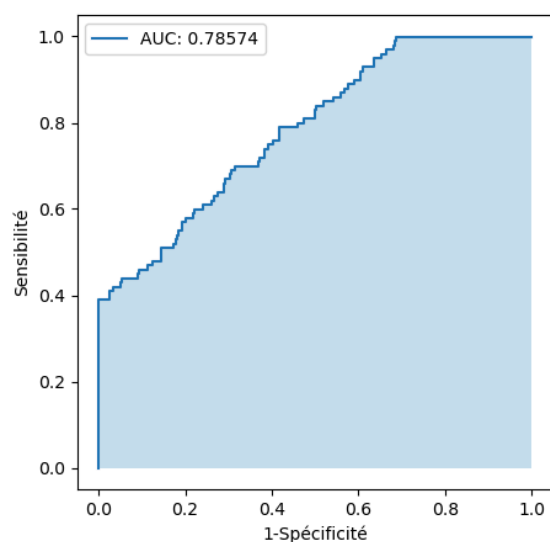
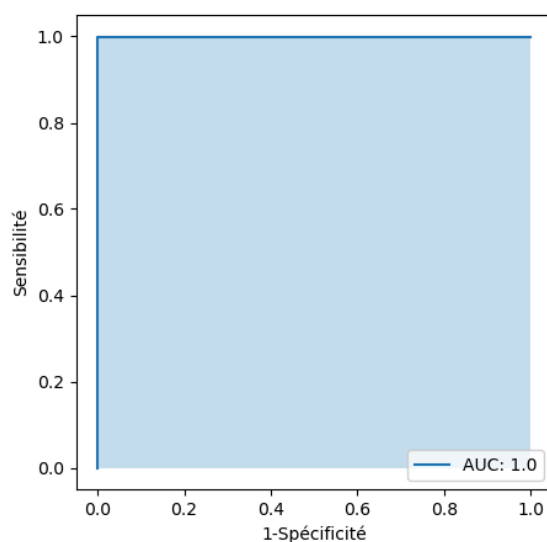
Courbe Précision/Sensibilité Exactement de la même manière que pour la courbe ROC, il est également possible de définir la courbe précision-sensibilité. À nouveau, on la synthétise en un seul nombre en extrayant de celle-ci l'aire sous la courbe. Cependant, pour cette métrique, il est courant de ne fournir qu'une approximation de cette aire suivant la méthode des rectangles simplifiant le calcul de l'intégrale. On parle alors de *moyenne de précision* (*Average Precision* en anglais, à comprendre dans le sens d'une moyenne sur la précision pondérée par la sensibilité) qui s'obtient suivant :

$$\text{AP} = \sum_n (\text{sensibilité}(s_n) - \text{sensibilité}(s_{n-1})) \cdot \text{precision}(s_n) \quad (3.75)$$

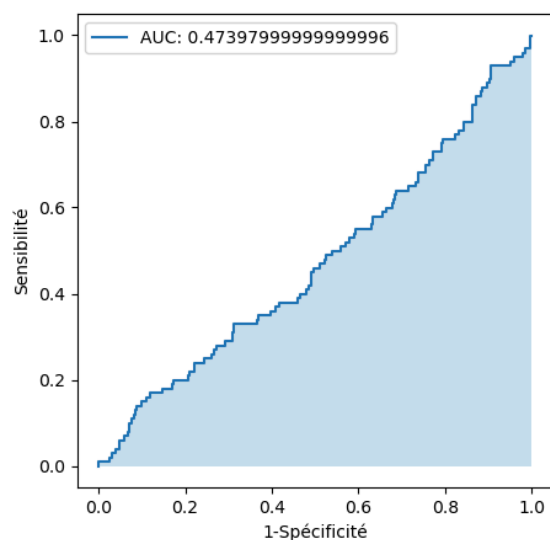
Par rapport à l'AUROC, cette mesure présente l'avantage de bien mieux refléter le déséquilibre de classes grâce à sa dépendance à la précision. Un modèle possédant une forte sensibilité mais une faible précision tend à trouver tous les éléments positifs, mais au prix d'en détecter trop (d'où un nombre très important de faux positifs) et inversement. Cette capacité à détecter ce déséquilibre en fait une métrique appréciée dans le cadre des mesures de segmentation, en particulier pour des structures de petites tailles. Néanmoins, pour cette tâche, il existe des métriques spécifiques appréciées par la communauté que nous présentons maintenant.

3.4.2 Segmentation sémantique

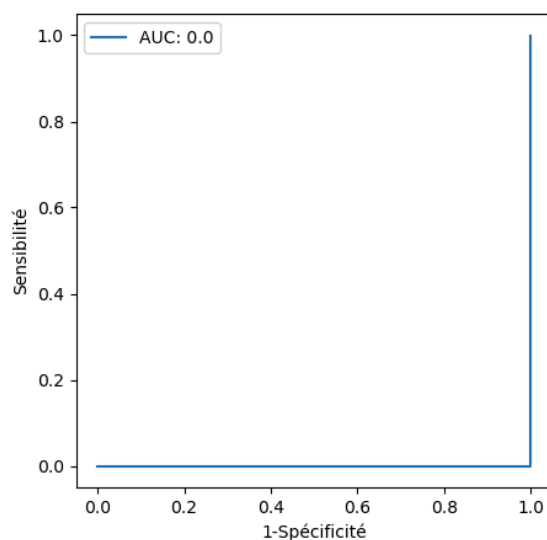
La segmentation peut toujours se ramener à un problème de classification, par conséquent, toutes les métriques précédentes s'appliquent aussi ici ; en gardant les mêmes considérations sur le choix judicieux de tel ou tel score en fonction des données (et notamment du déséquilibre de classes). Cependant, on attend aussi d'une métrique qu'elle possède un sens tangible et intuitif pour l'utilisateur. Cette intuition dépend des données et de la mesure : par exemple, si un modèle obtient une exactitude de 99.9% mais que la classe positive ne représente que 0.1% des pixels de l'image, cela ne dit rien de pertinent sur la performance du modèle (il peut n'avoir détecté aucun vrai positif). Dans ce cas de figure, l'exactitude se

(a) *Bon* modèle

(b) Modèle idéal



(c) Modèle faisant une prédiction aléatoire



(d) Modèle toujours faux

FIGURE 3.9 Exemples typiques de courbes ROC (et AUROC associés) pour différents types de modèles

révèle une mesure contre-intuitive du fait du déséquilibre des données. Une mesure qui prend en compte ce déséquilibre n'est pas nécessairement plus intuitive : si le coefficient de Kappa de Cohen κ d'un modèle est proche de 0, on saura de facto alors qu'il n'a obtenu aucun vrai

positif. Mais un Kappa de 0.3, 0.5 ou 0.7, bien qu'indiquant une hiérarchie des performances relatives, n'est pas beaucoup plus informatif sur la tendance du modèle dans l'absolu : fait-il de la sous ou de la sur-segmentation, manque-t-il une classe en particulier ?

C'est pour ces raisons qu'il est courant d'utiliser des mesures dites géométriques pour quantifier la performance d'un modèle de segmentation. Ces métriques se basent en général sur la superposition de structures géométriques, les rendant ainsi visuellement intuitives. Formellement, on considère les pixels de la vérité terrain et ceux segmentés comme étant deux ensembles X et Y . Pragmatiquement, sur des images, ces ensembles représentent bien des formes géométriques.

Indice de Jaccard Pour le calcul de l'indice de Jaccard, on définit alors le rapport de l'intersection sur l'union des deux ensembles :

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \quad (3.76)$$

Par analogie avec les métriques de classification, on peut exprimer $|X \cap Y| = \text{VP}$ et $|X| + |Y| - |X \cap Y| = \text{VP} + \text{FP} + \text{FN}$. On obtient alors l'interprétation :

$$J \cong \frac{\text{VP}}{\text{VP} + \text{FP} + \text{FN}} \quad (3.77)$$

Dans la littérature, ce score est plus souvent référé comme étant l'IoU (de l'anglais *Intersection-over-Union*) ou encore le mIoU (*mean-IoU*) pour son équivalent multi-classe.

Indice de Sørensen-Dice Le coefficient Dice (tel qu'il est le plus souvent référé dans la littérature) est également une métrique très utilisée. Il s'obtient par :

$$D(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \quad (3.78)$$

En reprenant l'analogie avec les notations de la matrice de confusion, on remarque que :

$$D \cong \frac{2 \text{VP}}{2 \text{VP} + \text{FP} + \text{FN}} \quad (3.79)$$

C'est-à-dire le score F1. L'intérêt de cette métrique est donc essentiellement de procurer une interprétation géométrique du dit score.

Les deux mesures (Dice et indice de Jaccard) sont en réalité très similaires et évoluent dans

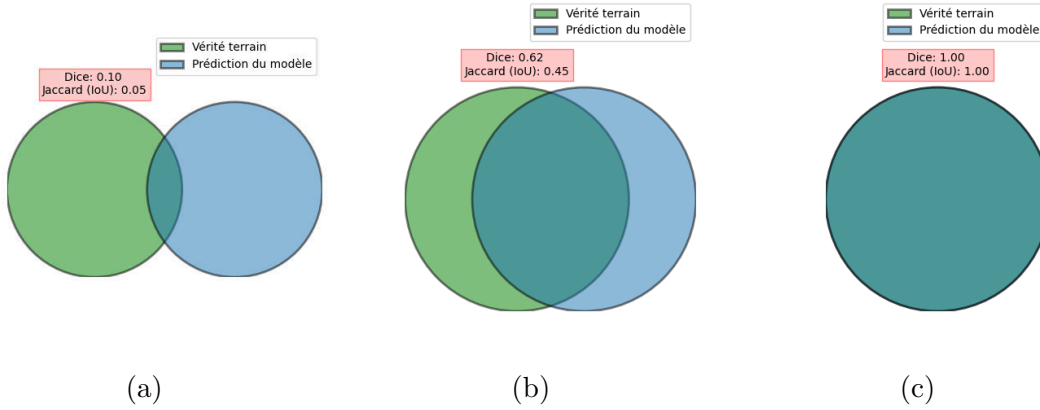


FIGURE 3.10 Évolution de l'indice de Jaccard et du coefficient de Sørensen-Dice sur trois cas de figure théoriques

le même sens. Il est d'ailleurs possible d'expliciter la relation qui les unit :

$$D = \frac{2 \cdot J}{J + 1} \quad (3.80)$$

L'étude de cette fonction nous renseigne notamment sur la double inégalité suivante :

$$D/2 \leq J \leq D \quad (3.81)$$

Ces résultats se retrouvent bien sur l'illustration qu'on donne des deux métriques sur la figure 3.10. En résumé, on y observe bien que par rapport à l'indice Jaccard, le coefficient de Dice accorde un plus gros poids aux éléments appartenant à l'intersection des ensembles, c'est-à-dire aux vrais positifs.

3.4.3 Un outil dynamique pour appréhender les métriques

Face à la multitude de métriques existantes, choisir celle(s) adaptée(s) au problème que l'on traite n'est pas trivial. Des travaux comme ceux de Maier-Hein et al. [124] fournissent bien une méthodologie de sélection, mais une fois le choix fait, l'interprétation d'une métrique donnée n'est pas forcément évidente pour autant. C'est particulièrement vrai pour certaines applications où la notion de performance reste souvent vue comme relativement subjective : ainsi, pour la segmentation sémantique, un utilisateur aguerri peut en un coup d'œil comparer la prédiction à une vérité terrain. Or, l'appréciation qualitative ne rejoint pas forcément l'évaluation quantitative que fournit une métrique. Par exemple, sur la figure 3.10b, quali-

tativement, on pourrait croire à un indice de Jaccard relativement élevé; il n'est pourtant « que » de 0.45. Pour interpréter un score, il est également nécessaire de se former à sa lecture sur des cas concrets, au delà de la compréhension du mode de calcul. À cette fin, nous avons développé une plateforme en accès ouvert sur internet visant à se familiariser sur les différentes métriques à travers des cas concrets manipulables. L'outil est conçu à des fins pédagogiques et se concentre sur la segmentation sémantique (une capture d'écran est fournie sur la figure 3.11). Depuis sa publication initiale, deux modules ont été ajoutés, pour traiter des métriques de classification générale et celles concernant les séquences temporelles⁶.



FIGURE 3.11 Outil interactif de visualisation des métriques de segmentation. L'utilisateur se voit présenter deux canevas sur lesquels il peut dessiner (à gauche, la vérité terrain et à droite, la « prédiction »). Les métriques sont actualisées en fonction des tracés pour illustrer leur évolution. Plusieurs scénarios pré-établis sont configurés pour illustrer les forces et les faiblesses de différentes métriques implémentées.

6. L'outil est accessible à l'adresse suivante (sur ordinateur uniquement) : <https://clementpla.github.io/SegmentationMetricTutorial/welcome>

3.5 Applications aux pathologies rétinienne

Jusqu'à présent, cette revue de littérature a fourni un inventaire des concepts théoriques utilisés de près ou de loin dans les travaux de recherche menés au cours de cette thèse. Or, celle-ci s'inscrit dans un contexte médical : si les notions théoriques sur la conception d'architecture, de généralisation, d'interprétabilité et autres ont été manipulées durant nos travaux, c'est dans le contexte d'une recherche appliquée à l'ophtalmologie. Or, ce champ plus spécifique a également vu de nombreuses innovations se développer. La suite de cette revue de littérature propose donc de les couvrir et d'aborder plus précisément les différentes recherches et applications de l'apprentissage machine déployées autour de l'imagerie rétinienne.

3.5.1 Segmentation des structures rétinienne dans l'imagerie de fond d'œil

Dans la pratique clinique, le diagnostic de la rétine se fait sur la base de l'analyse de biomarqueurs structurels dans l'imagerie de fond d'œil ou de l'OCT. Ainsi, la présence de certaines lésions ainsi que leur distribution est un indicateur primordial de l'état d'avancement de certaines maladies. Mais l'analyse des structures normales de la rétine peut également révéler une anormalité, que ça soit dans le processus d'acquisition ou par rapport à l'état de santé du patient. Par exemple, le nombre d'embranchements visibles dans l'arbre vasculaire peut servir d'indicateur de qualité de l'acquisition. Cette section propose un survol des méthodes proposées pour la segmentation de telles structures mais par esprit de synthèse, seules les approches les plus récentes sont introduites. L'accent est mis sur le fond d'œil, notamment car il s'agit de la modalité la plus facile d'accès et la plus proche de nos travaux.

À l'instar de l'ensemble des publications récentes en vision par ordinateur, l'écrasante majorité des travaux de ces dernières années sur la segmentation sémantique de l'imagerie de fond d'œil s'appuient sur des réseaux de neurones convolutifs. La prédominance de ce modèle unique ne doit pas masquer l'existence en pratique de très nombreuses contributions originales visant à l'ajuster aux spécificités de la tâche. En général, les types de réseaux proposés suivent toujours les grands principes des architectures fondatrices présentées en section 3.1.4, tandis que d'autres recherches se concentrent sur l'optimisation en elle-même, par exemple via l'introduction de fonctions de coûts adaptées aux contraintes des données.

Les premières publications dédiées à la segmentation de lésions rétinienne ont en commun de ne se concentrer que sur un seul type de lésions. En 2016, Van Grinsven et al. [130] proposent un simple réseau (type LeNet) pour la segmentation des hémorragies rétinienne, reposant sur un principe de classification par patch (présence ou absence). Du fait de leur inefficience, ce type de modèle est aujourd'hui complètement dépassé en faveur des équivalents totalement

convolutifs (et donc faisant de la classification par pixel). Chudzik et al. [131] proposent une architecture utilisant le modèle de l'*Inception Block* (voir figure 3.3) visant à segmenter les exsudats dans des patches d'images de fond d'œil. Les performances sont évaluées sur la base de données E-Ophtha et mettent particulièrement en exergue l'appartenance du transfert d'apprentissage (y compris via pré-apprentissage sur des petites bases de données). Les mêmes auteurs proposeront plus tard [132] un nouveau modèle hybride composé d'un réseau convolutif utilisé d'une part pour fournir une carte de segmentation et d'autre part l'extraction d'un vecteur de caractéristiques duquel est dérivé un score de proximité entre l'image d'entrée et sa plus proche voisine dans la base d'apprentissage.

À notre connaissance, le travail que nous avons mené dans le passé (initialement dans [133] puis prolongé dans [134]) est la première publication portant sur la segmentation multi-classes (donc sémantique) par réseaux de neurones des lésions visibles au fond d'œil. Pour ce travail, l'architecture utilisée s'inspire du U-Net (présenté dans la section 3.1.4) adaptée pour posséder un double décodeur. Chacun d'eux est chargé de la segmentation respective des lésions sombres (hémorragies, micro-anévrismes) et des lésions claires (drusen, exsudats durs et doux). Cette méthodologie propose également une approche originale de l'exploitation de données faiblement labellisées (c'est à dire pour lesquelles on ne dispose que du diagnostic et non des cartes d'annotations de lésions). L'article, bien que démontrant globalement des performances élevées en terme de segmentation, conclue sur la nécessité de constituer des bases de données plus conséquentes. Par ailleurs, la discrimination en deux classes seulement des lésions (sombres ou claires) est critiquable pour son manque de finesse. Il n'empêche que ce travail fut riche en enseignements sur les difficultés spécifiques à la tâche. Celles-ci n'auront cesse de motiver de nouvelles améliorations en terme d'architecture de réseaux de neurones dans les travaux subséquents. On peut citer notamment :

- La difficulté de segmenter des structures ayant une grande variabilité de forme, de texture et de taille. Cette disparité pose de nombreux problèmes, y compris aux annotateurs humains.
- La présence de lésions potentiellement extrêmement petites (quelques pixels de large) au sein des images de très hautes résolutions. Outre la difficulté d'extraire un signal contrasté de ces lésions parmi le bruit de fond de la rétine, il existe un très fort déséquilibre de classe auquel les techniques d'apprentissage statistiques sont très sensibles.
- L'absence de standard approprié pour évaluer la qualité d'une segmentation sémantique sur des structures aussi diverses.
- Un nombre généralement très faible de données d'entraînement comparativement à d'autres domaines d'applications des réseaux de neurones.

Plusieurs travaux ont depuis été publiés, prenant en compte un plus grand nombre de classes

dans leur modèle et traitant de ces problématiques. Par ailleurs, plusieurs équipes ont également joint leurs données aux publications, offrant à la communauté de nouvelles bases de données publiques avec annotations réalisées par des médecins. Les travaux de Xu et al. [135] étendent le principe d’une architecture U-Net à une problématique multi-classes. En l’occurrence, tout comme pour la majorité des travaux qui seront détaillés dans cette section, quatre classes sont considérées :

1. Les exsudats (parfois également appelés *exsudats durs*)
2. Les hémorragies
3. Les micro-anévrismes
4. Les exsudats doux ou nodules cotonneux (CWS)

Plusieurs modifications sont proposées pour améliorer l’architecture initiale : d’une part, chaque couche de l’encodeur est enrichie d’un module *MSFF* (*Multiscale Feature Fusion*), librement inspiré de l’*Inception Block* décrit en section 3.3. De plus, la fusion des caractéristiques entre le décodeur et l’encodeur s’appuie sur un modèle d’attention, là encore librement inspiré du module *Squeeze & Excitation* (voir section 3.1.1). On observe que l’utilisation d’une architecture multi-échelle est souvent présentée comme efficace pour gérer la variabilité de taille entre les lésions. Guo et al. en intègrent le concept dans leur réseau, surnommé L-Seg [136], reprenant une architecture de type VGG16 (voir la section 3.1.1) de laquelle sont extraites des cartes de caractéristiques à cinq échelles différentes. Celles-ci sont ensuite fusionnées indépendamment pour chacune des quatre classes considérées.

Yan et al. [137] introduisent un modèle intégrant à la fois le contexte global en moyenne résolution et local en haute résolution. Pour cela, deux réseaux sont entraînés en parallèle, l’un sur l’image entière, l’autre sur un patch d’image extrait aléatoirement. Les caractéristiques extraites de ces deux modèles sont fusionnées par concaténation avant d’être transformées en carte de segmentation. Si les performances rapportées de ce double modèle sont très élevées, on peut noter que la reproductibilité de cet article est rendue difficile par son manque de précision (par exemple, les auteurs n’indiquent pas explicitement s’ils ont entraîné leur modèle sur chaque lésion simultanément ou individuellement).

Guo et al. proposent une architecture similaire, nommée CARNet [138], visant à raffiner une segmentation globale (faite sur l’image entière) avec une segmentation par patches en haute résolution. Mais à la différence de Yan et al. [137], la fusion des caractéristiques est faite à chaque couche de l’architecture de décodage, par un mécanisme nettement plus complexe reposant sur une pondération par attention. Cette complexité architecturale, couplée à un code non publié et une méthodologie manquant de détails rend à nouveau la reproductibilité de cet article difficile.

Bases de données Une des difficultés de la segmentation sémantique du fond d’œil est liée à la diversité des images, due à des facteurs aussi divers que l’ethnicité des patients, leur état de santé (par exemple, une cataracte tend à assombrir l’image), le matériel d’acquisition etc. À cela s’ajoute que les lésions observables sur la rétine sont de formes, couleurs et textures très variables, ce qui mène à une annotation manuelle particulièrement laborieuse et coûteuse. Ces deux constats (diversité des images, difficulté d’annotations) rendent très précieuses les bases de données publiées qui sont au fondement d’un très grand nombre d’articles de recherche. Cinq bases de données se démarquent, notamment en raison de leur inter-compatibilité supposée (c’est-à-dire qu’elles possèdent les mêmes classes d’annotations) :

1. Dataset for Diabetic Retinopathy (DDR), publiée par Li et al. [139]. Dans leur étude, les auteurs évaluent les performances de réseaux existants sur la classification de la rétinopathie diabétique et la segmentation des lésions, obtenant des performances modérées sur cette dernière tâche.
2. Fine-Grained Annotated Diabetic Retinopathy (FGADR), publiée par Zhou et al. [140]. Il s’agit de la plus grande base de données publique annotées de cette liste.
3. IDRiD, publiée à l’occasion d’une compétition organisée par Porwal et al. [141].
4. Retinal Lesions, publiée par Wei et al. [142]. Les auteurs proposent par ailleurs une architecture multitâches, capable de simultanément segmenter les lésions et classifier la rétinopathie diabétique.
5. Messidor, publiée par Decencière et al. [143], cette base est à l’origine dédiée à la gradation de la rétinopathie diabétique. Au cours de ce doctorat, un panel d’experts ont ré-annoté 200 images issues de cette base pour un ensemble de biomarqueurs (lésions et structures normales).

Une comparaison détaillée de ces différentes bases est proposée dans le chapitre 4. Notons simplement qu’elles se distinguent les unes des autres par leur taille, leur qualité et la finesse des annotations réalisées par les experts.

3.5.2 Classification de l’imagerie rétinienne

La première publication portant sur le diagnostic automatisé de l’état d’un fond d’œil remonte à 1976, détaillant une étude de Yamamoto et Yokouchi [144]. Leur système, semi-automatique, vise à classifier les croisements veines/artères. Le processus n’étant pas entièrement automatisé, un clinicien est tout d’abord chargé d’identifier la zone de croisement. L’algorithme va

alors extraire les contours des vaisseaux, puis en dériver un diagnostic par la mesure du ratio des diamètres de la veine et de l'artère. Si cette méthode algorithmique est excessivement simple (selon les standards contemporains), il est intéressant de noter que certains points abordés restent très actuels ; en particulier le traitement de la variabilité des couleurs existante entre plusieurs images de fond d'œil, variabilité qui à l'époque tirait son origine de la pellicule choisie pour l'acquisition, mais qui aujourd'hui dépend principalement du type de capteurs ou de l'ethnicité du patient. La difficulté que présente cette variabilité au sein des images (liée à l'acquisition, mais aussi anatomique) explique l'important travail de recherche sur cette modalité depuis cette première étude. Encore aujourd'hui, cet effort de recherche ne s'est pas relâché. Et si l'OCT n'existait pas encore à l'époque, il s'agit aujourd'hui d'un instrument communément utilisé. La suite de cette section traitera par conséquent à la fois des travaux menés sur le fond d'œil et ceux portant sur l'OCT. Elle sera divisée par ailleurs en deux axes. Le premier aborde la question du diagnostic rétinien par approche globale. Dans cette catégorie est incluse toute méthode qui ne s'appuie pas sur la reconnaissance explicite d'une structure (saine ou pathologique) anatomique. Une approche de classification de maladies reposant sur des caractéristiques mathématiques (texture, fréquences, contours...) est considérée comme globale. À l'inverse, la section suivante abordera le diagnostic par approche locale, c'est à dire par la segmentation préalable de marqueurs anatomiques pour établir un diagnostic. Dans le cas de la rétinopathie diabétique par exemple, cela passera par la détection de lésions ; pour le glaucome, par la segmentation complète du nerf optique, etc. A ce titre, cette approche sera également qualifiée d'ascendante, car allant de l'échelle locale (c'est-à-dire au niveau du pixel) pour remonter à une prédiction globale (au niveau de l'image). La différence fondamentale avec l'approche globale réside par conséquent dans le fait d'associer localement un sens à chaque pixel (ou région de pixels) de l'image, créant de fait une relation causale : *Tel(s) marqueur(s) identifié(s) détermine(nt) tel diagnostic.*

Ces deux catégories permettent de distinguer les modèles globaux, purement statistiques, des modèles locaux, reposant sur une connaissance biologique/médicale a-priori. Pour autant, il existe des modèles inversant l'approche globale (partant de modèles purement statistiques pour essayer de retrouver des marqueurs locaux causaux), de telle sorte que cette distinction devra être nuancée. En particulier, une courte discussion sera consacrée aux modèles interprétatifs. En substance, ceux-ci illustrent bien ce mécanisme d'inversion : suite à une classification globale, le modèle est conçu pour proposer une explication locale : c'est-à-dire des marqueurs cliniquement interprétables.

3.5.3 Approche statistique de la classification de pathologies

L'approche statistique (ou globale) de la classification de pathologies repose sur le présupposé que non seulement les données sont intrinsèquement statistiquement classables grâce à leur seul contenu mais surtout qu'il n'est pas nécessaire de baser cette classification à une pré-identification de certaines structures locales spécifiques via une expertise externe (autrement dit via un savoir a priori). L'algorithme développe cette expertise en construisant lui-même la représentation statistique adaptée à chaque classe de diagnostic. Les plus modernes, basés sur des réseaux de neurones pour la plupart, excellent généralement à cette tâche. Cependant, l'approche globale se heurte à deux obstacles principaux :

- Pour apprendre une représentation de la distribution de chaque classe que l'on souhaite identifier, on estime nécessaire de disposer d'un nombre important de données, représentatif de tous les cas possibles. La capacité de généralisation sur un cas clinique non-prévu durant l'apprentissage est incertaine.
- L'acceptabilité clinique d'une telle solution, fut-elle validée quantitativement, est limitée par l'absence d'interprétation accompagnant la prédiction.

Cette section parcourt une brève revue générale de différentes solutions proposées sur les deux modalités que sont le fond d'œil et l'OCT.

Méthodes classiques pour l'imagerie de fond d'œil

Traditionnellement (c'est-à-dire avant les réseaux de neurones) la classification de maladies repose sur une méthodologie en deux étapes : la première consiste à extraire des caractéristiques discriminantes de l'image et la seconde à classer ces caractéristiques dans les catégories diagnostiques (typiquement sain/malade dans le cas binaire, sain/grade 1/grade 2/.../grade n pour l'estimation de la sévérité ou bien sain/maladie 1/maladie 2/.../maladie n dans le cas multi-classes). Les caractéristiques peuvent être extraites automatiquement (pour les approches reposant sur un apprentissage de bout en bout) soit en utilisant des modèles mathématiques traduisant l'image en une somme d'informations jugées les plus pertinentes (c'est à dire discriminantes) pour l'étape de classification. Cependant, pour bien distinguer cette dernière de l'approche locale (décrite dans la section 3.5.4), spécifions d'emblée que ces caractéristiques ne correspondent pas à des structures anatomiques définies. De ce fait, l'étape d'extraction de caractéristiques ne peut et n'est pas évaluée par rapport à une base de données annotées par des experts. Par conséquent, seul le résultat final de classification est évalué par rapport à la vérité terrain.⁷

7. Cette définition peut mener à des cas ambigus, où des caractéristiques anatomiques sont extraites (par exemple, la segmentation des microanévrismes) mais où les performances de segmentation ne sont pas

En dehors des réseaux de neurones convolutifs (permettant un apprentissage de bout en bout incluant l'extraction de caractéristiques), les principales méthodes d'extraction de caractéristiques reposent sur l'analyse spectrale et/ou texturale. On citera Guo et al. [145], proposant une méthode de classification de la sévérité de la cataracte, entraînée et évaluée sur 445 images de fond d'œil. Les caractéristiques employées sont les coefficients caractéristiques de la transformation en ondelettes de Haar de l'image, combinés à une transformation en cosinus discrets de celle-ci. La classification se fait par analyse factorielle discriminante de Fisher. Acharyau et al. [146] proposent d'utiliser les moments statistiques spectraux d'ordre élevé (le moment d'ordre 0 étant la moyenne, celui d'ordre 1 la variance) pour caractériser une image, selon une technique issue de la théorie du traitement du signal. Le vecteur résultat est utilisé pour la classification des quatre stades de la RD en utilisant une Machine à Vecteurs Supports (*Support Vector Machine*) (SVM). Une approche similaire est proposée à nouveau par Acharyau et al. [147], proposant une classification basée sur la classification de textures du fond d'œil. Deux types de caractéristiques de textures sont utilisées. La première découle de l'utilisation de la matrice de co-occurrence de l'image (de laquelle sont dérivées des mesures de contraste, d'homogénéité, d'entropie...), la seconde utilise des statistiques d'ordre plus élevées. A nouveau, un classifieur SVM est utilisé pour distinguer les stades de la rétinopathie. Chowriappa et al. [148] proposent une approche similaire pour la classification de la maculopathie diabétique (plus exactement, la classification des œdèmes maculaires comme cliniquement significatifs ou non). Cependant, les informations texturales extraites sont différentes (dimension fractale, motifs binaires locaux (*Local Binary Pattern*) (LBP), ondelettes...). Par ailleurs, l'étape finale de classification repose sur un ensemble de quatre classificateurs (deux étant des variantes du classificateur Bayésien, le troisième un arbre de décision et le dernier une variante des SVM).

L'information texturale a également été employée pour d'autres maladies que la rétinopathie diabétique. Mookiah et al. [149] proposent une variante des LBP, nommée *Local Configuration Pattern* pour distinguer les différentes classes de la DMLA. 22 caractéristiques texturales sont extraites par image et classées par un ensemble de classificateurs (Arbre de décision, k-Plus Proches Voisins (k-PPV), classification Bayésienne, Réseau de Neurones Probabilistes et SVM). Les auteurs insistent particulièrement sur la sélection des caractéristiques et du classifieur optimaux. En particulier, l'effet de la taille du vecteur de caractéristiques est précisément quantifié.

quantifiées, c'est à dire une approche locale sans comparaison avec une annotation clinique au niveau du pixel. Auquel cas, seule la classification est réellement évaluée, mais ces travaux seront cependant considérés comme appartenant à l'approche locale.

Classification du fond d’œil par réseaux de neurones

Ces différents exemples donnent un aperçu des approches globales employées sur le fond d’œil en utilisant des modèles pré-définies d’extraction de caractéristiques. Bien que notre revue ne soit pas exhaustive, il est notable que le nombre de publications sur le sujet est relativement restreint par rapport à l’approche locale et l’approche globale par apprentissage profond. À l’inverse, pour celle-ci, la recherche menées ces dernières années est extrêmement prolifique. Devant l’impressionnante quantité d’articles publiés sur le sujet⁸, plusieurs revues de littérature (y compris parues la même année) ont été proposées (par exemple [150–156] pour n’en citer que quelques-unes). Elles ont en grande partie constitué notre point d’entrée pour dresser l’inventaire qui suit.

La vaste majorité des travaux que nous couvrons se base sur des variations d’architectures de réseaux de neurones convolutifs. Pour la rétinopathie diabétique, Gulshan et al. [157] ont démontré la capacité d’un tel réseau à distinguer le stade sain des stades modérés et plus avancés de la rétinopathie diabétique. Le principal intérêt de cette étude est d’avoir mené l’évaluation de leur méthodes sur un jeu important de données (11 171 images provenant de 5871 patients). En obtenant une sensibilité de 97.5% et une spécificité de 93.4%, les auteurs ont démontré la viabilité clinique des réseaux de neurones, ces performances étant comparables voire supérieures à celles d’un ophtalmologiste. Gargeya et al. [158] ont confirmé ces résultats par une approche très semblable sur le même jeu de données ; la différence étant l’utilisation de la validation-croisée et une architecture de classification légèrement différente. Pratt et al. [159] ont mené une étude similaire, en entraînant cette fois leur algorithme à faire de la classification du grade de la RD (et non juste du dépistage), suivant l’échelle ETDRS. Les résultats obtenus sont en revanche relativement peu concluants : la spécificité obtenue est de 95% pour une sensibilité de 30% ; on note notamment une incapacité complète du réseau à distinguer la classe « RD légère », systématiquement confondue avec les classes « Saine » et « Modérée ». Ting et al. [160] proposent une approche basée sur un ensemble de réseaux (entraînés en parallèle) visant à détecter plusieurs maladies conjointement. Les quatre prédictions produites concernent le dépistage de la DMLA, d’un potentiel glaucome, de la RD de stade modérée (ou plus) et de la RD évoluée au stade de menace pour la vision du patient. Pour l’ensemble de ces classifications, les résultats obtenus sont légèrement en deçà de ceux des cliniciens expérimentés. Le système résultant est validé sur un ensemble de dix bases de données différentes provenant de communautés hétérogènes. Pour le dépistage de la RD (premier stade), les performances varient pour la sensibilité entre 97.1% et 100 %

8. Une recherche sur le référenceur Google Scholar avec la combinaison de mots-clés *deep learning, fundus, classification, diabetic retinopathy* couvrant la période 2019-2023 aboutit à 1610 articles publiés avec révision par les paires (13 300 tout type de publications confondu).

pour une spécificité allant de 73.3% à 92.2%. Pour un stade plus avancée de la RD menaçant la vision, le système obtient un score de détection d'AUC sensibilité/spécificité de 95.8%. Pour la suspicion de glaucome, ce même score est de 94.2% et de 93.2% pour la DMLA. Li et al. [161] proposent également un réseau de neurones pour détecter la présence de marqueurs du glaucome dans le fond d'œil, évaluée sur 8000 images extraites d'une base de données de 48 116 images. La sensibilité (respectivement spécificité) obtenue est de 95.6% (resp. 92 %), pour une AUC de 98.6%. Une analyse des résultats obtenus par le réseau de neurones montre que la principale source d'erreurs de classification (faux négatifs et faux positifs) provient de l'existence d'une autre pathologie présente dans le fond d'œil, en particulier une forte myopie, une RD et/ou une DMLA.

Plus récemment, une étude similaire a été publiée par Sahlsten et al. [162]. Les auteurs abordent le problème en concevant une architecture permettant la classification des cinq classes de RD ainsi que de quatre classes d'OMD, en utilisant une base de données plus restreinte (mais quand même composée de 28512 images d'entraînement). La principale contribution de ce travail est d'évaluer l'importance de la résolution des images d'entrées dans la classification des maladies. Les auteurs proposent une étude comparant les résolutions allant de 256×256 jusqu'à 2095×2095 , cette dernière étant exceptionnellement élevée pour un CNN. L'étude met en évidence l'importance de ces très hautes résolutions pour l'obtention de performances compétitives sur la détection des deux maladies. Une étude de même ampleur a également été menée sur la classification de la DMLA par Grassmann et al. [163]. Le système entraîné se compose de six architectures de réseaux de neurones conventionnelles (VGGNet, Inception, ...), dont les prédictions respectives sont agrégées par une forêt d'arbres décisionnels. En pratique, les auteurs ont distingué différents stades de la maladie liée à l'âge de la rétine (neuf stades différents pour autant de classes, selon l'échelle de gradation de l'étude proposée par Davis et al. [164]), trois classes de progression des derniers grades de la DMLA et une classe dédiée aux images considérées non-évaluables (c'est-à-dire de mauvaise qualité). Le modèle complet est entraîné sur 120656 images et évalué sur 5555 images issues d'une base indépendante. Les performances varient assez fortement en fonction des tranches d'âges et des données utilisées, mais le réseau obtient en moyenne une exactitude de 85.7% (top-2, signifiant que parmi les deux classes prédites obtenant les scores les plus élevés se trouve la classe effective de l'image).

La tendance actuelle visant à substituer les modèles auto-attentionnels aux CNNs n'a pas épargné la classification du fond d'œil, avec un nombre relativement important de nouveaux articles adaptant le ViT à la détection de la rétinopathie diabétique. La très récente revue de littérature proposée par Muchuchuti et al. [165] offre un aperçu des améliorations que cette architecture peut apporter, que l'on retrouve d'ores et déjà appliquées à plusieurs reprises à

la gradation de la rétinopathie diabétique [166–169]. Les performances rapportées sont généralement au niveau de celles des architectures CNNs équivalentes en nombre de paramètres ou d’opérations. En revanche, le mécanisme d’attention est très souvent plébiscité pour sa capacité à générer des prédictions facilement interprétables.

Tomographie par Cohérence Optique

Comparativement au fond d’œil, l’état de l’art sur la classification automatique des images OCT est relativement réduit, ce qui s’explique par le fait que la technologie soit nettement plus récente et cliniquement plus difficile à manipuler. En parallèle, là où les images de fond d’œil sont désormais à portée de main de tout chercheur, le nombre de données OCT accessibles publiquement est considérablement plus restreint, ce qui peut constituer un obstacle à la recherche transdisciplinaire.

De plus, la difficulté à interpréter les données OCT a limité le nombre de travaux portant sur l’approche globale. En effet, les données étant généralement plus bruitées et moins contrastées que le fond d’œil tout en étant plus volumineuses, l’accent est généralement porté sur la détection de marqueurs cliniquement interprétables (en particulier la segmentation des différentes couches rétinienne). Mais il existe cependant quelques exemples d’approches reposant sur une interprétation globale. Ces travaux suivent globalement la même trajectoire chronologique que ceux traitant du fond d’œil présentés dans la section précédente. Là où les premières études dérivait d’abord des vecteurs caractéristiques extraites de l’image via des modèles empiriques paramétrés avant d’en faire la classification dans un second temps, la tendance actuelle porte de plus-en-plus à une classification automatique de bout-en-bout. L’approche de Srinivasan et al. [170] appartient à cette première catégorie empirique. Pour faire la classification entre des patients sains, atteints d’OMD ou de DMLA les auteurs emploient un SVM pour classer des vecteurs caractéristiques extraits de chaque B-scans. Chaque vecteur est obtenu en calculant l’Histogramme des Gradients Orientés (*Histogram of Oriented Gradients*) (HOG). Les *B-scan* ont été préalablement pré-traités pour corriger le bruit Speckle ainsi que la courbe de la rétine (qui est une surface tridimensionnelle sphérique) projetée vers un plan. 45 patients ont contribué à l’étude. D’autres travaux remplacent l’HOG par un vecteur basé sur les LBP (éventuellement sous différentes variantes). Lemaitre et al. [171] les utilisent dans la détection des OMD sur une cohorte de 32 patients. Dans ce cas, les LBP sont calculés en prenant en compte le volume. La méthode obtient une spécificité de 93.7% et une sensibilité de 81.2%. Albarrak et al. [172] utilisent une combinaison des LBP et des HOG pour la détection de la DMLA. Là encore, les caractéristiques sont extraites par sous-volumes et non par tranches. 140 acquisitions OCT sont utilisées et le système atteint une spécificité de 90.5% pour une sensibilité de 92.4%. L’ensemble de ces études utilisent les SVM comme

classifieurs, ceux-ci étant connus pour être capables de généraliser de façon robuste y compris avec un petit nombre de données d’entraînements.

A contrario de ces approches exploitant des vecteurs caractéristiques explicites, Lee et al. [173] sont les premiers à expérimenter une approche par CNN (architecture VGG) pour le dépistage de DMLA. L’étude est menée sur 9285 patients, menant à un total de 43 328 acquisitions OCT centrées sur la macula. Le réseau entraîné obtient une sensibilité de 85.41% pour une spécificité de 93.82% sur 20% de cette base de données préservées pour l’évaluation. Karri et al. [174] étendent ces travaux à une classification tri-classes, permettant en plus de la DMLA de distinguer les OMD. N’ayant accès qu’à 45 patients, les auteurs s’emploient à démontrer l’efficacité de l’apprentissage par transfert, en récupérant un réseau pré-entraîné à la classification d’images naturelles (non médicales).

La liste des travaux se basant sur l’apprentissage profond et les réseaux convolutifs pour la classification ne cesse de s’agrandir, élargissant le champs d’application et confirmant l’efficacité globale de ces algorithmes. Cependant, la conséquence néfaste de cette tendance est qu’il devient peu aisé de distinguer une contribution propre. A la lecture de la revue de littérature qui vient d’être donnée, on observe un point marquant : Un certain nombre de travaux et en particulier ceux reposant sur l’apprentissage par CNN, se limitent à l’étude de chaque *B-scan* individuellement traité comme une image. Or l’OCT donne accès à une représentation 3D, et même si des contraintes matérielles et techniques limitent les traitements numériques réalisables, on s’attend à ce que la prise en compte des informations volumiques renforcent la capacité prédictive d’un modèle. En particulier, l’utilisation d’un réseau convolutif tri-dimensionnel paraît appropriée à la tâche de classification.

3.5.4 Du local au global : approche ascendante de la classification des pathologies par reconnaissance préalable des biomarqueurs anormaux

Dans cette section seront présentés différents travaux reposant sur une décomposition de la détection de la maladie en deux étapes. La première est une étape de segmentation (ou de détection), visant à obtenir les contours (ou respectivement de trouver la région) d’une structure pré-définie et dont la connaissance permet de guider un diagnostic. La seconde étape consiste à analyser ces résultats de segmentation, d’en extraire des caractéristiques spécifiques et, suivant certaines règles cliniques ou statistiques, s’en servir pour établir le diagnostic d’une maladie donnée.

Imagerie de fond d'œil par approche locale

Rétinopathie diabétique L'essentiel des méthodes utilisées pour la détection et/ou la gradation de la RD repose sur une pré-segmentation des lésions, notamment les micro-anévrismes, et sur des heuristiques cliniques (nombre de lésions, répartition par rapport à la macula, type de lésions...) et nous renvoyons à la section 3.5.1 pour une revue de littérature plus détaillée sur le sujet. Certaines approches ne reposent cependant pas sur la détection de lésions. C'est le cas de celle de Welikala et al. [175], qui proposent d'identifier le stade prolifératif de la RD en se basant sur la carte du réseau vasculaire (obtenu par segmentation) dont est extrait un ensemble de caractéristiques. La segmentation des vaisseaux est obtenue par une variante du détecteur de ligne multi-échelle proposé par Nguyen et al. [176]. Ces caractéristiques propres au réseau vasculaire traduisent des informations topologiques et/ou morphologiques (nombre de pixels appartenant à un vaisseau, nombre de segments, densité, orientations...). Le vecteur caractéristique obtenu est utilisé pour distinguer les images contenant des néo-vaisseaux et celles n'en contenant pas. Une approche similaire est employée par Christodoulidis et al. [177], qui proposent une caractérisation des petits vaisseaux segmentés par une approche par tenseur. L'approche permet ainsi d'identifier localement les néo-vascularisations.

De cet exemple, on perçoit la pertinence d'un système complet pour la gradation de la rétinopathie diabétique, c'est à dire incluant à la fois une segmentation des lésions, mais également de l'arbre vasculaire.

Imagerie OCT par approche locale

La détection de maladies par approche locale dans les images OCT est généralement initiée par une étape de segmentation des couches rétiniennes, éventuellement des lésions, d'extraction de caractéristiques (notamment d'épaisseur) et d'une dernière étape de classification.

Rétinopathie diabétique Cette approche est employée par Eltanboly et al. [178], qui segmente 12 couches rétiniennes distinctes (ajoutant les zones d'interdigitation, ellipsoïde et myoïde et la membrane limitante externe à la liste dressée en section 2.1.1). La segmentation est obtenue par un ensemble de règles relativement complexes, basées sur la transformation de l'image d'entrée en une carte de probabilités (en utilisant la théorie des champs aléatoires de Gibbs-Markov). Une fois les couches segmentées, des caractéristiques d'épaisseur, courbure et réflexivité en sont extraites puis classées par un réseau de neurones.

Lésions Novosel et al. [179] sont les premiers à proposer une méthode de segmentation de lésions couplée à la segmentation des couches rétinienne et obtenant les mêmes performances que l'état de l'art propre à chacune de ces deux tâches. Par ailleurs, la méthode ne repose pas sur l'apprentissage machine, la rendant, selon les auteurs, plus facile à améliorer. Chaque couche (les lésions étant considérées comme une couche) est segmentée suivant une équation de propagation. Fang et al. [180] proposent une méthode de classification des lésions maculaires (et non de maladie), également par une première étape de segmentation. La classification se fait par B-scan, parmi quatre classes (Drusen, OMD, CNV ou Saine). Le mécanisme présenté repose sur la notion d'attention (même si le terme est légèrement galvaudé ici, l'attention étant contrainte et non apprise). Un premier réseau est entraîné à identifier une région contenant une lésion (quel que soit son type). La prédiction de ce premier réseau est fournie en entrée d'un second, qui se chargera de classer le B-Scan, en étant conditionné par la pré-segmentation du premier réseau. Une idée similaire est proposée par Huang et al. [181], à travers un double CNN de segmentation/classification. Un premier réseau pré-entraîné est chargé de segmenter les différentes couches rétinienne ainsi que les masses fluides. Pour ce rôle, les auteurs ont choisi le ReLayNet, librement accessible et utilisable, proposé par Roy et al. [182]. De la carte de segmentation obtenue sont extraites la couche située entre la RPE et la membrane interne supérieure d'une part et d'autre part la couche située entre la RPE et la membrane de Bruch. Chacune des deux couches est fournie à un second jeu de CNNs, qui prédisent la présence de Drusen, d'OMD et de CNV.

3.6 Discussion

On peut tirer plusieurs enseignements de cette revue de littérature, en particulier de son versant consacré à l'ophtalmologie :

- Pour le fond d'œil, l'application principale est la détection et gradation de la rétinopathie diabétique, en raison notamment de l'existence de plusieurs grandes bases de données publiques consacrées à cette maladie. Il est déjà établi que les modèles existants offrent, dans certaines conditions, des performances comparables à un expert humain, le potentiel d'améliorations est donc limité. De plus, la plupart des gains de performances obtenus récemment font appel à des données privées et inaccessibles.
- En revanche, la segmentation des structures rétinienne dans le fond d'œil continue à stimuler la recherche, en témoigne notamment la multiplication récente des bases de données accessibles publiquement contenant des annotations manuelles réalisées par des experts. Celles-ci sont unanimement reconnues comme longues et coûteuses à produire ; leur publication est donc une opportunité de recherche. Cependant, la

diversité des bases interrogent sur leur inter-comptabilité et la capacité d'un modèle à apprendre malgré leur hétérogénéité. Nous n'avons pas identifié de travaux qui s'attèlent directement à cette question.

- Si la classification est un problème largement couvert dans la littérature, elle souffre souvent de dépendre d'algorithmes peu transparents dans leur prise de décision. Une des raisons est qu'il n'existe pas de biais inductifs contraignant réellement un modèle de type CNN à s'appuyer sur les structures locales cliniquement identifiées comme des lésions pour faire son diagnostic. Même les travaux combinant segmentation et classification ne garantissent pas que l'utilisation de la première serve à la seconde. De manière générale, la pertinence de la segmentation est un aspect assez peu couvert dans la littérature : il est admis que les médecins se basent sur des lésions pour faire le diagnostic, il est donc attendu que l'algorithme fasse de même, cependant dans la pratique, cette hypothèse est rarement confrontée.
- Concernant l'OCT, les bases de données publiques sont moins nombreuses et généralement plus petites, à l'exception de celle proposée par Kermany et al. [183], qui se limite à quatre classes (normal, drusens, CNV et OMD). Les B-scans qui la composent sont tous acquis avec un Spectralis de la gamme *Heidelberg Engineering*. En général, les différents algorithmes publiés sur cette base font preuve déjà d'excellentes performances, mais la généralisation sur d'autres appareils d'acquisitions est rarement couverte. Par ailleurs, là aussi, l'interprétabilité des modèles est encore rarement mise en avant.

Ces différents éléments ont façonné les objectifs de recherche que nous avons présentés en section 1.2. Les prochains chapitres rentrent dans le vif des expériences que nous avons réalisées pour les atteindre.

CHAPITRE 4 SEGMENTATION DES LÉSIONS RÉTINIENNES DANS L'IMAGERIE DE FOND D'ŒIL

Pour le médecin, l'identification et la quantification des biomarqueurs visibles dans l'image de fond d'œil est un préalable au diagnostic des maladies affectant potentiellement la rétine. En effet, les règles diagnostiques d'un grand nombre d'entre elles se basent sur la reconnaissance d'une ou plusieurs structures anatomiques et de leur analyse (topologie, nombre, texture, etc.). Parce qu'elle illustre parfaitement ce concept, prenons la rétinopathie diabétique comme exemple : chaque grade de la maladie est caractérisé par un ensemble de marqueurs (essentiellement des lésions) dont l'évolution (en taille, nombre et type) est révélatrice de la progression de la gravité de la maladie [184]. En pratique, cette identification est loin d'être aisée, y compris pour des cliniciens expérimentés. Qui plus est, le recours de plus en plus fréquent aux centres de téléépistages prive l'examineur d'informations contextuelles lui permettant de résoudre certaines ambiguïtés. Nous illustrons dans la section 4.0.1 suivante plusieurs des difficultés pouvant tromper y compris l'œil expert. À ce titre, confier la tâche à un algorithme est perçue comme une solution potentielle. L'identification des frontières des structures locales et la reconnaissance de celles-ci dans une image sont appelées segmentation sémantique. Ce chapitre traite de la question de l'obtention des cartes de segmentation pour chaque grande classe de lésions existante dans le fond d'œil, mais aussi de l'exploitation de ces segments une fois ces cartes obtenues.

À travers notre revue de littérature sur le sujet (section 3.5.1) se dessine l'image d'un champ de recherche actif et massivement tourné vers la conception et l'optimisation de nouvelles architectures de réseaux de neurones. Pour subvenir aux besoins de cette recherche, de plus en plus d'équipes ont mis en accès libre leurs données pour un usage académique, ce qui représente souvent l'équivalent de plusieurs centaines d'heures de travail d'annotations réalisées par des experts. En revanche, rares (sinon inexistantes) sont les travaux qui explicitent leur protocole d'annotations de manière détaillée et reproductible. Cela mène à des annotations interbases très hétérogènes, bien que représentant supposément les mêmes structures. À cela s'ajoute que chaque base de données représente une certaine population et a été acquise dans des conditions spécifiques (caméra, condition d'éclairage, acquisition mydriatique ou non...). Or, étant donnée la taille relativement réduite de ces bases, il apparaît clairement qu'une représentation exhaustive de toutes les spécificités du fond d'œil à travers une seule base confine à la gageure. Pourtant, il n'existe à notre connaissance aucun travail cherchant à combiner ces bases de données (que ce soit pour entraîner ou évaluer un modèle) ou même simplement à les caractériser respectivement. La littérature tend globalement à décorréliser

les travaux sur la segmentation en elle-même et les questions afférentes à la généralisation. L'étude présentée dans ce chapitre propose donc de combler cette absence d'analyse comparative. Plusieurs objectifs de recherche vont venir préciser et élargir largement l'ambition initiale de ce travail et sont détaillés dans la section 4.1. Pour délimiter notre étude, le choix a été fait de ne traiter que des cinq bases de données récentes les plus importantes. Par ordre de taille (c'est-à-dire de nombre d'images), celles-ci sont :

1. Indian Diabetic Retinopathy Image Dataset (IDRiD) [141].
2. MESSIDOR [143]
3. Dataset for Diabetic Retinopathy (DDR) [139]
4. RET-LES [142]
5. Fine-Grained Annotated Diabetic Retinopathy (FGADR) [140]

Une description détaillée de ces bases est proposée dans la section 4.0.2. Ces jeux de données ont l'avantage d'avoir en commun certaines annotations de lésions spécifiques, sur lesquelles porte notre travail :

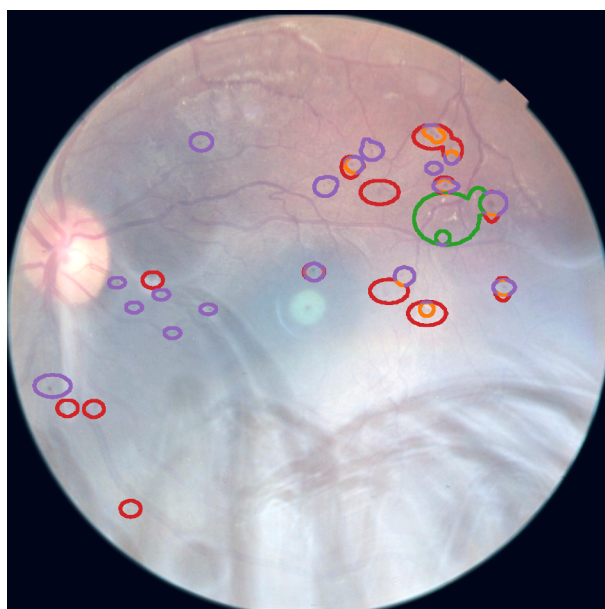
1. Exsudats (EX) (également qualifiés d'*exsudats durs*)
2. Hemorragies (HEM)
3. Micro-anévrismes (MA)
4. Cotton-Wool Spot (CWS)

Notre étude s'arrime donc à la tendance de la majorité des articles publiés qui se restreint à ces quatre lésions. Notons qu'il existe des travaux étendant leur portée à un plus grand nombre de lésions ; en particulier pour la segmentation des ischémies et des néovascularisations [142], qui sont de toute première importance pour la classification des stades avancés de la rétinopathie [184]. Mais notre étude se voulant comparative, il était nécessaire de se limiter au dénominateur commun en matière de classes traitées.

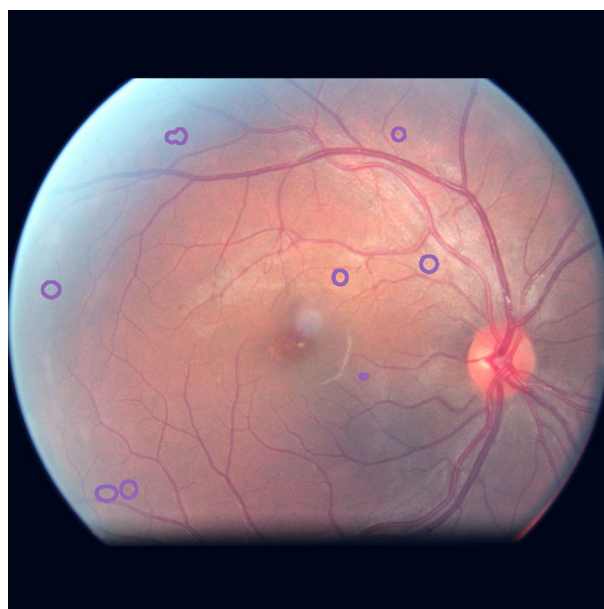
4.0.1 Difficultés liées à la segmentation des lésions

Il est notoire - et les efforts continus de recherche en attestent- que la segmentation des lésions rétinienues n'est pas une tâche aisée ; même en faisant abstraction des difficultés liées à la disparité entre les bases de données. Nous recensons sur la figure 4.1 un certain nombre de cas illustrant la complexité du problème, toutes les images étant issues de la même base RET-LES.

Nous avons également eu l'occasion de vérifier cette difficulté liée à l'identification des lésions en conduisant une mini-étude auprès de trois experts de la rétine. Nous leur avons demandé



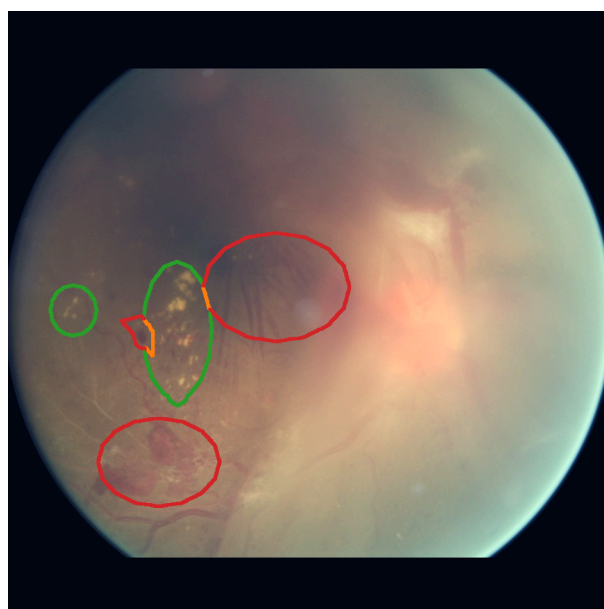
(a) Artefacts visibles dans la région inférieure (cils)



(b) Surexposition lumineuse qui sature le champ temporal de la rétine



(c) Sous-exposition globale compliquant la lecture de la périphérie temporale supérieure de la rétine



(d) Faible contraste dans le champ nasal dû à une pathologie (détachement rétinien ou cataracte)

FIGURE 4.1 Images issues de la base de données RET-LES. L'annotations des lésions est indiquée en surimpression (rouge : hémorragies, vert : exsudats, violet : microanévrismes, orange : nodules cotonneux).

de faire la gradation de la rétinopathie diabétique sur 200 images ainsi que l'identification de la présence d'œdème maculaire, suivant une échelle classique de 0 à 4. Il en ressort notamment douze images (6%), pour lesquelles il existe un désaccord total entre les trois observateurs (c'est-à-dire qu'aucune paire ne s'accorde). Après révision une à une des images en session, le désaccord s'avéra porter sur l'identification de certaines structures en tant que lésions de type microanévrismes. Or, sur cinq des douze images, une poussière sur le capteur de la caméra, visible dans la région maculaire, a trompé deux des trois observateurs sur la présence d'un microanévrisme. Le dernier observateur a constaté la présence de cette anomalie en comparant les positions (toujours identiques) de ce faux anévrisme sur ces 5 images consécutivement. Cette anecdote témoigne du type de difficulté existant dans la segmentation rétinienne et du défi qu'elle représente, y compris pour des experts humains.

4.0.2 Variabilité interbases

Les cinq bases de données que nous analysons font état d'importantes interdisparités à deux niveaux :

- en termes d'images en elles-mêmes ; les facteurs de variabilité étant l'apparence de l'image (terme générique regroupant la colorimétrie, le contraste, l'exposition...), la résolution, la qualité d'acquisition et la diversité des symptômes des patients présents dans la base.
- en termes d'annotations réalisées par les médecins ; cette fois, ce sont la finesse (c'est-à-dire la précision aux frontières) des annotations, leur quantité et le nombre de classes différentes annotées qui font cette diversité.

Bien entendu, à cela s'ajoute que les bases de données sont de taille diverse. Le tableau 4.1 fournit le détail de la composition de chaque base.

La variabilité statistique des annotations dépend grandement du protocole d'annotation, sur lequel nous n'avons en général que peu d'informations. Or, la variabilité qu'il induit est très importante : nous en proposons une estimation quantifiable sur la figure 4.2. La finesse des annotations d'une base comme IDRiD suggère une assistance par ordinateur pour faciliter un marquage extrêmement détaillé. Nous pouvons attester que cette stratégie est utilisée sur MESSIDOR, le processus d'annotations ayant été réalisé dans le cadre de ce doctorat en collaboration avec des cliniciens ayant vérifié et corrigé les annotations prérealisées automatiquement par un modèle. Inversement, sur RET-LES, les annotations suggèrent un tracé à partir de formes géométriques simples prédéfinies. Entre ces deux extrêmes, les styles varient : comme on peut le voir sur la colonne correspondant à FGADR, celle-ci présente à la fois des annotations aux frontières fines (trois premières lignes) et plus grossières (quatrième ligne).

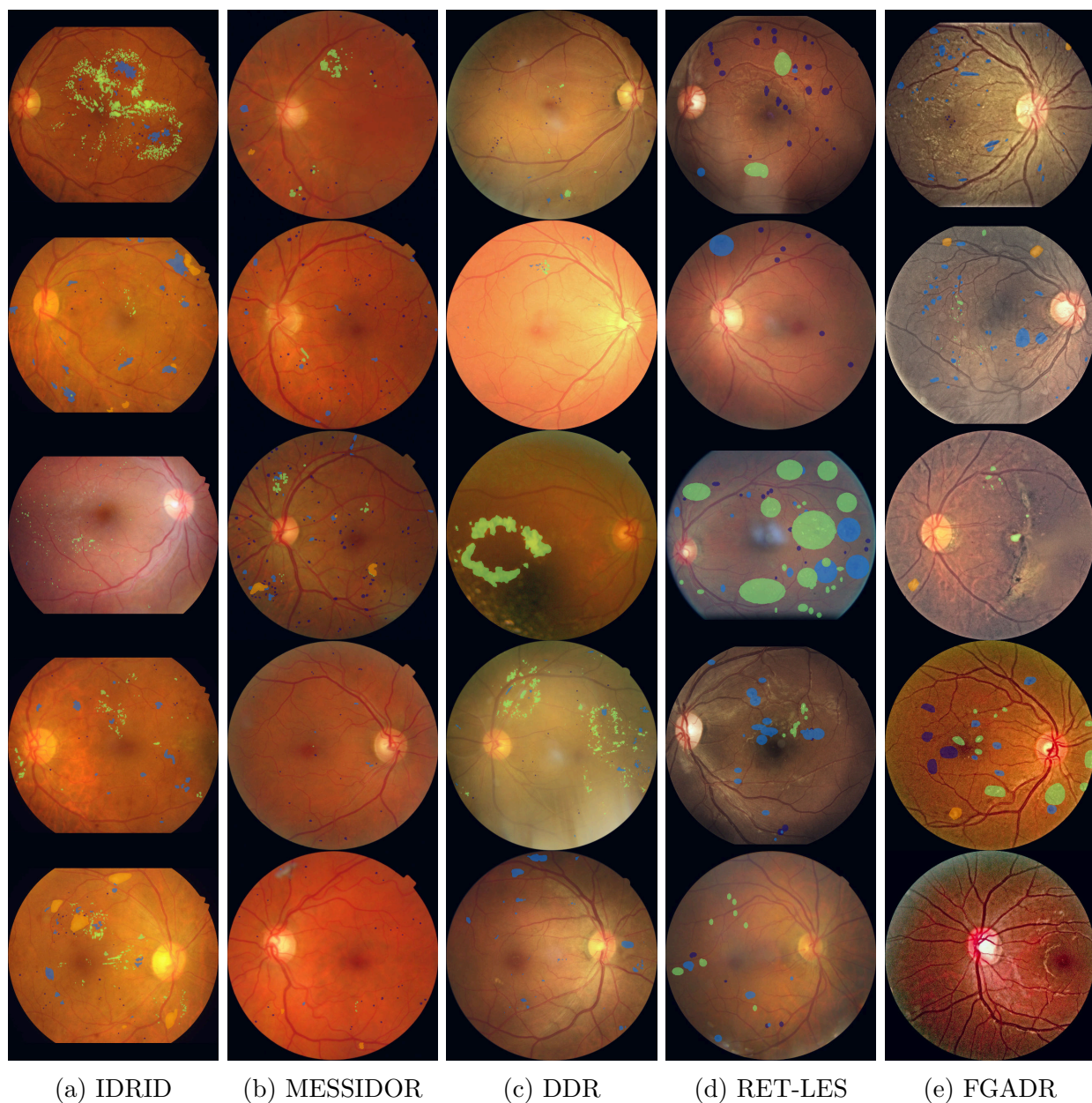


FIGURE 4.2 Images présentées avec leurs annotations manuelles et organisées par colonne suivant leur base d'appartenance. Que ce soit dans les couleurs, le découpage de la région d'intérêt ou les annotations, on constate une grande variabilité inter et intrabase.

TABLEAU 4.1 Composition de chacune des bases de données utilisées. Hémorragie intravitréenne (HIV), Hémorragie intra-rétinienne (HIR), Hémorragie pré-rétinienne (HPR), Anomalie Microvasculaire Intra-rétinienne *Intraretinal Microvascular Abnormality* (IrMA), Néovascularisation (NV)

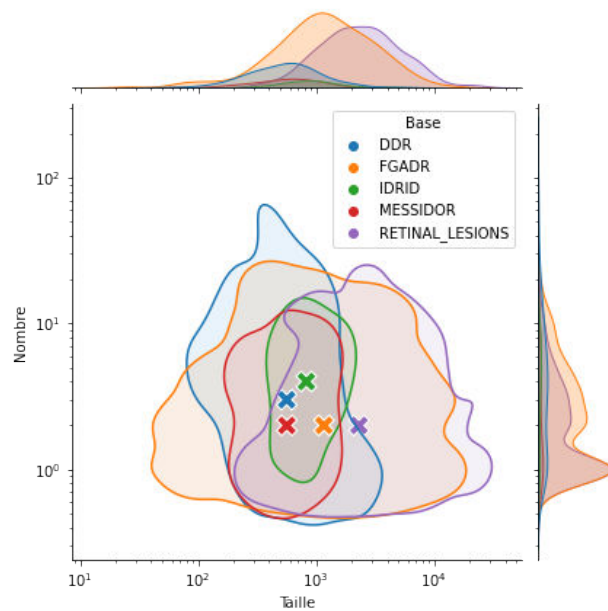
Bases	Nombre d'images	Résolution	Classes annotées
IDRiD	81	2848×4288	MA, HEM, CWS, EX
MESSIDOR	200	1500×1500	MA, HEM, CWS, EX, Drusen, Macula, Disque Optique, Coupe Optique, HPR, NV, Vaisseaux Sanguins
DDR	757	1934×1956	MA, HEM, CWS, EX
RET-LES	1593	896×896	MA, HEM, CWS, EX, NV, HIV, HIR, Prolifération Fibreuse
FGADR	1842	1280×1280	MA, HEM, CWS, EX, NV, IRMA

Il est possible de se faire une idée plus quantifiée du style d'annotation propre à chaque base. Pour cela, nous suggérons une analyse statistique de la distribution jointe du nombre de lésions et de leur taille moyenne par image. Bien entendu, ces deux variables dépendent du diagnostic de la maladie sous-jacente, mais ici toutes les images considérées sont pathologiques (à des degrés de sévérité divers). Abstraction faite de la maladie, la seconde dépendance entre ces variables vient donc du style d'annotation. En particulier, deux régimes se distinguent : pour une annotation de granularité fine, l'expert aura tendance à marquer de nombreux éléments de manière précise et détaillée. Inversement, une annotation plus grossière aura tendance à fusionner les structures locales voisines en des structures plus grosses et moins nombreuses. La figure 4.3 montre ces distributions jointes pour les quatre lésions et pour chacune des cinq bases de données. Elle met bien en évidence l'observation qualitative : IDRiD se distingue par un grand nombre d'éléments segmentés de petite taille, suivi généralement de DDR et MESSIDOR. RET-LES est systématiquement la moins fine, toutes lésions confondues (celles-ci étant moins nombreuses et de plus grande taille). Enfin, tel que visible sur la figure 4.2, FGADR se retrouve à cheval entre ces deux tendances, présentant des images possédant de nombreuses lésions de petites tailles et d'autres en possédant peu, mais de grandes tailles.

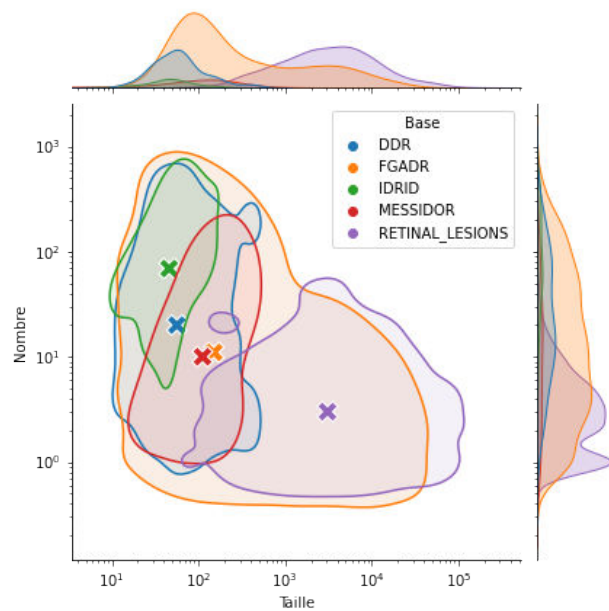
4.1 Problématique

Considérant les éléments relatifs aux deux sections précédentes (c'est-à-dire les difficultés liées à la segmentation et la grande variabilité interbase existante), plusieurs éléments de problématique se posent :

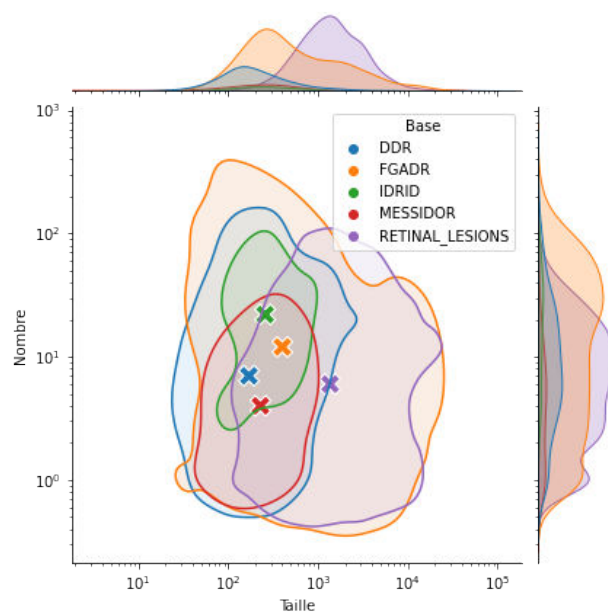
- quelles sont les capacités de généralisation d'un modèle donné sur diverses bases de



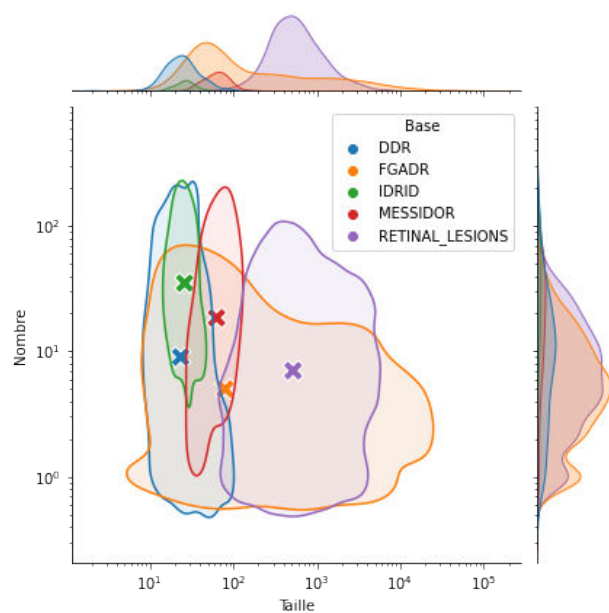
(a) Distribution des CWS



(b) Distribution des EX



(c) Distribution des HEM



(d) Distribution des MA

FIGURE 4.3 Représentation des distributions jointes de la taille moyenne et du nombre de lésions par images pour chaque classe. Les croix représentent les centroïdes de chaque distribution. Pour pouvoir représenter les distributions sur un même graphique, il a fallu choisir une échelle logarithmique, il y a donc plusieurs ordres de grandeur séparant les centroïdes. En comparant les bases de données, on met ainsi en évidence la disparité des styles d'annotations entre chacune.

données ?

- corollairement, que peut-on conclure des performances d'un modèle à partir d'une seule base ?
- quel impact le style d'annotations propre à chaque base va-t-il avoir sur le diagnostic de la pathologie ?

À travers les deux premières questions, nous dressons un premier objectif de recherche, qui est de concevoir un modèle capable de segmenter plusieurs lésions simultanément. Nous ajoutons deux contraintes techniques sur le choix du modèle : celui-ci doit pouvoir être facilement reproductible (donc sans ambiguïté d'implémentation) et ses performances doivent a minima égaler les modèles de l'état de l'art. Ce modèle nous servira de base pour évaluer les problématiques de généralisation.

Notre second objectif de recherche porte sur l'évaluation de la capacité de généralisation empirique du modèle sur différentes bases afin de dresser le tableau des compatibilités inter-bases.

Une question de recherche qui vient se greffer à cet objectif est celle de la pertinence des métriques utilisées pour l'évaluation de la qualité de la segmentation. Pour répondre à ce questionnement, un autre volet d'analyse sera traité, qui décrit la relation entre la qualité de la segmentation et de la détection des lésions.

En vue d'uniformiser les performances interbases, il est assez intuitif de vouloir réduire la variabilité entre les images présentées par un effet de prétraitement. Nous proposons d'explorer l'effet qu'induit un tel type de traitement sur le modèle.

Nous nous attendons à trouver des disparités au sein des modèles de segmentation dépendamment de leur base d'entraînement. Notre objectif principal est à la fois d'en étudier les conséquences en matière de segmentation sur des images de test, mais aussi de proposer une méthode pour contrôler le comportement des modèles en fonction du besoin clinique. On notera cependant que celui-ci n'est pas clairement défini : dans quelle mesure ces disparités de segmentation sont-elles cliniquement pertinentes ou dommageables ?

Pour clarifier cette question, nous proposons dans un dernier temps d'étudier la capacité d'un modèle à classer une image à partir des lésions identifiées en elle.

4.2 Méthodologie

Notre méthodologie s'articule autour de quatre volets, dont chacun fait l'objet d'une section dans la partie qui suit :

1. La standardisation via des traitements algorithmiques simples des images d'entrée.
2. La conception d'une architecture de segmentation, incluant sa comparaison avec l'état

de l'art.

3. La mise en place d'une stratégie permettant d'uniformiser les performances d'un modèle à travers plusieurs distributions de données.
4. Le développement d'un modèle de classification permettant l'évaluation de la pertinence de la segmentation à des fins diagnostiques.

4.2.1 Normalisation des données

Les images utilisées varient en résolution et en apparence. La phase de prétraitement vise donc à corriger ces deux aspects et est présentée dans cette section. La standardisation des résolutions est une nécessité technique, notamment du fait que les modèles ne sont théoriquement pas invariants en échelle ; elle est donc menée pour chacune de nos expériences. En revanche, l'intérêt de l'homogénéisation de l'apparence est plus discutable. Pour alléger ce chapitre et à la vue de résultats mitigés, nous le mentionnons brièvement dans la section 4.2.1, mais le détail des résultats est reporté en annexe C.1.

Standardisation des résolutions

La résolution d'image varie dépendamment de la base. Afin de standardiser le format d'entrée de nos modèles, notre protocole inclut les étapes suivantes :

1. La région d'intérêt (rétine) est extraite. Pour cela, on utilise un simple effet de seuillage sur le canal rouge de l'image.
2. L'image est ensuite découpée pour éliminer les bordures noires et ainsi être centrées sur la région d'intérêt.
3. Si l'image résultante n'est pas de format carré, on utilise du *padding* (complétion des bords par des zéros) sur la dimension la plus petite pour lui conférer ce format.
4. L'image est ensuite redimensionnée au standard que nous avons adopté de 1536×1536 .

On peut noter que cette résolution est supérieure à celle des images fournies dans RET-LES et FGADR. Pour ces deux bases, les images apparaîtront comme légèrement moins détaillées (la sous-résolution initiale pouvant agir comme un filtre passe-bas). En pratique, nous avons quantifié cet effet, en utilisant des images de très hautes résolutions (type IDRID), filtrées par un passe-bas pour simuler une sous-résolution. Nous n'avons pas observé de baisse de performances particulièrement significatives sur les CNNs entraînés dépendamment de ce filtrage, ce qui semble indiquer la validité du choix d'une résolution standard intermédiaire parmi toutes celles présentes dans les bases. Précisons tout de même que la résolution choisie reste exceptionnellement élevée : dans la littérature, la plupart des modèles sont entraînés sur

des images de taille 512×512 . Plusieurs travaux font cependant état d'une corrélation entre qualité du dépistage et résolution des images, il paraît donc pertinent d'adopter un standard élevé pour celle-ci.

Algorithme de prétraitement

La nécessité et l'impact d'un prétraitement des données -visant à homogénéiser l'apparence des images indépendamment de leur origine- sont encore très débattus dans la littérature. Nous avons mené nos propres expériences en la matière. Le protocole et les résultats détaillés sont fournis en annexe C.1, mais en voici les conclusions résumées : quel que soit l'angle d'analyse choisi (par lésions ou par bases de données), les gains potentiels obtenus par un traitement visant à normaliser l'apparence des images semblent limités. En revanche, il existe des cas pour lesquels ils entraînent des situations de faillite complète de convergence du modèle, mais nous n'avons pas exploré les raisons expliquant ces configurations défectueuses. Pour la suite du chapitre, nous considérons donc les images dans leur apparence initiale, sans algorithme de prétraitement.

4.2.2 Conception d'une architecture polyvalente

Le choix d'une architecture de réseau de neurones est sujet à de nombreuses considérations sur les données et le budget calculatoire disponible. De nombreux travaux sur la segmentation du fond d'œil mettent en avant (voir la revue de littérature en section 3.5.1) la grande variabilité de taille des lésions et proposent des modèles combinant une approche locale et globale pour la prendre en compte. Mais en pratique, notre revue de littérature ne fait pas ressortir une architecture indiscutablement plus performante que les autres, trop de variables (nombre de paramètres, résolution d'images, prétraitement, itérations d'entraînement) étant interdépendantes sans qu'il y ait de consensus sur leur importance respective. De plus, les variations de performances sont souvent relativement minimales, qui plus est sur des métriques bien précises et potentiellement sensibles à des microvariations (par exemple dues aux frontières de segmentation, par nature ambiguës). On observe également souvent une absence de standard dans le rapport fourni des résultats, que ce soit sur le choix de la métrique ou son mode de calcul. Heureusement, des concours comme IDRiD ont permis de corriger cet aspect dans une certaine mesure, standardisant l'utilisation de certaines métriques et facilitant ainsi les comparaisons. Enfin, pour l'essentiel de la littérature, on note que seules les performances de segmentation sont rapportées et non celles de détection. Or, aucun travail ne semble s'interroger sur la pertinence d'une segmentation fine d'un point de vue diagnostique par rapport à une simple détection : autrement dit, la précision sur les frontières de segmentation n'a,

pour l’heure, pas démontré sa pertinence clinique.

En dressant cet inventaire des freins à la comparaison à l’état de l’art, on dessine en creux une critique d’une tendance assez fréquente : celle de voir apparaître de nouvelles architectures introduites sans prendre en compte l’existant et la modularité des réseaux de neurones existants. Nous arguons qu’en l’absence de consensus sur ce qui fait la performance pratique d’un modèle, il est relativement superflu de se concentrer sur la construction de nouvelles architectures spécifiques au fond d’œil.

Au contraire, nous suggérons d’utiliser la modularité des réseaux de neurones, qui nous permet d’obtenir une pléthore de combinaisons parmi les architectures existantes. Ces modèles ont l’avantage d’avoir été largement étudiés et éprouvés dans la littérature sur des bases de données généralistes autrement plus grandes et diverses. Par ailleurs, ces modèles sont plus facilement reproductibles, y compris dans leur version préentraînée, car nous les avons extraits de bibliothèques publiques connues.

Pour appuyer notre propos et afin de sélectionner l’architecture qui nous servira à conduire notre étude de généralisation, nous proposons d’explorer les forces de cette modularité architecturale, en combinant deux blocs. Toutes les architectures testées fonctionnent sur un même modèle de type :

$$y = \mathcal{D}(\mathcal{E}(x)) \quad (4.1)$$

où x, y sont respectivement les entrées et les sorties du modèle, $\mathcal{E}$ représente un module d’encodage et $\mathcal{D}$ de décodage. L’encodeur produit une ou plusieurs cartes de caractéristiques à différentes résolutions et le décodeur a la tâche de les combiner en une unique carte de segmentation. À partir de cette description simple, nous pouvons piocher dans la littérature pour concevoir séparément nos modules d’encodage et de décodage en respectant des structures d’architectures existantes. Le tableau 4.2 reporte l’ensemble des combinaisons essayées. Le fonctionnement des différents modules est décrit dans la section 3.1.4. Certains modules ne sont pas compatibles entre eux (ce qui explique les cases vides du tableau). En considérant ceux-ci, on aboutit à un total de 27 architectures comparées. Cela nous permet de couvrir une large gamme de modèles, à la taille et à la complexité très variables. Notons que tous les modèles ne sont pas convolutifs, en particulier, les encodeurs MIT B2/4 se basent sur l’architecture du SegFormer de Xie et al. [185] (utilisant des réseaux autoattentionnels).

TABLEAU 4.2 Architectures utilisées dans notre étude comparative, accompagnées de leur taille (paramètres et FLOPs)

Architecture	Encodeur	# Paramètres	# FLOPs
UNet	ResNet 18	14.33M	21.82G
	ResNet 34	24.44M	31.49G
	SE ResNet50	35.05M	41.67G
	ResNest - 50	34.45M	49.57G
	EfficientNet-B7	67.1M	39.2G
	MobileNetV3	3.59M	10.43G
	MIT B2	27.48M	31.6G
	MIT B4	64.12M	64.92G
UNet++	ResNet 18	15.97M	64.18G
	ResNet 34	26.08M	73.85G
	SE ResNet50	51.52M	0.23T
	ResNest - 50	50.91M	0.24T
	EfficientNet-B7	68.16M	75.32G
	MobileNetV3	3.71M	13.97G
	MIT B2	-	-
	MIT B4	-	-
FPN	ResNet 18	13.05M	17.85G
	ResNet 34	23.16M	27.53G
	SE ResNet50	28.65M	30.15G
	ResNest - 50	28.04M	38.05G
	EfficientNet-B7	65.67M	35.27G
	MobileNetV3	2.72M	8.29G
	MIT B2	26.08M	28.7G
	MIT B4	62.73M	62.03G
DeepLab V3+	ResNet 18	12.33M	18.33G
	ResNet 34	22.44M	31.63G
	SE ResNet50	29.21M	35.97G
	ResNest - 50	-	-
	EfficientNet-B7	65.11M	60.21G
	MobileNetV3	2.16M	2.98G
	MIT B2	-	-
	MIT B4	-	-

4.3 Résultats expérimentaux

4.3.1 Performances comparées entre les architectures de segmentation

Afin de sélectionner un modèle parmi ceux présentés dans la section 4.2.2, un protocole expérimental standardisé a été mis en place. Celui-ci s’inspire du concours IDRiD (dont est issue la base de même nom). Sur les 81 images de la base, 54 sont utilisées pour l’apprentissage (dont une fraction pour la validation), les 27 restantes pour l’évaluation du modèle. La métrique utilisée est l’Aire sous la Courbe (*Area Under the Curve*) -AUC- Précision/Sensibilité calculée sur 11 seuils uniformément répartis pour chaque lésion séparément. Pour nos entraînements, seules les données IDRiD sont utilisées, en revanche, tous les encodeurs ont été préentraînés sur ImageNet. Les modèles sont entraînés avec un pas d’apprentissage de 0.001 et avec une dégradation de la pondération (*weight decay*) de 0.00001 sur 1500 époques. Précisons qu’une époque correspond à un parcours de la base d’apprentissage. Comme les $\mathcal{B}^{(i)}$ varient en taille, les $\mathcal{M}[\mathcal{B}^{(i)}]$ ne sont donc pas tous entraînés avec le même nombre d’itérations totales ($= N_{\text{époques}} \times |\mathcal{B}^{(i)}|$). En revanche, chaque image est bien utilisée le même nombre de fois pour chaque entraînement. La fonction de coût utilisée est la généralisation différentiable du Dice introduite par Sudre et al. [186]. Différentes augmentations de données sont utilisées à la volée (modifications de contraste, luminosité, rotation, miroir...). Par ailleurs, lors de l’apprentissage, les images sont découpées en « patches » de taille 512×512 . Les performances obtenues par les modèles sont rapportées dans le tableau 4.3, qui offre une comparaison entre nos entraînements et les résultats rapportés dans la littérature. Toutes les combinaisons encodeur/décodeur mentionnées dans la section méthodologie (Tableau 4.2) n’apparaissent pas ; cela s’explique du fait que notre processus de sélection s’appuie sur l’algorithme d’optimisation *Hyperband* [187], proposé par Li et al. Cet algorithme permet de mettre fin de manière prématurée aux entraînements (en l’occurrence une des combinaisons encodeur/décodeur) non prometteurs. Pour cela, les performances sont suivies sur un ensemble de validation au cours de l’entraînement (on utilise le *mIoU* comme métrique). Un budget est accordé à chaque entraînement (nombre d’itérations d’apprentissage à réaliser, temps maximum etc.). Ce budget est réactualisé en fonction des configurations d’entraînement les plus prometteuses. Celles qui ne le sont pas (c’est à dire qui n’intègre pas le top-k des meilleurs *mIoU* obtenus jusqu’à présent) sont stoppées. Les autres voient leur budget augmenter et l’entraînement se poursuivre. Cette technique assure une économie significative de ressources lors d’une exploration de configurations. Le tableau 4.3 ne présente que les résultats des modèles ayant été entraînés en atteignant leur budget maximal.

Les configurations potentielles ont été choisies pour couvrir un ensemble de modèles de tailles diverses, utilisant les multiples techniques introduites dans notre état de l’art sur les architec-

tures de réseaux de neurones (section 3.5.1). Cet éventail couvre également des architectures mixtes, intégrant les récents modèles autoattentifs de type *Transformer* et *Segformer*.

4.3.2 Étude empirique de la généralisation interbases

La notion de généralisation est au cœur de la recherche sur la théorie de l'apprentissage statistique dont nous avons esquissé les contours dans la section 3.2. En revanche, nous abordons une approche résolument empirique, qui se détache des fondements théoriques.

Analyse quantitative des résultats

Pour faciliter la lecture de notre analyse, nous introduisons les conventions de notation suivantes :

- On introduit une indexation sur les différentes bases de données. Pour chaque base i , les ensembles d'entraînement et d'évaluation sont référés en tant que $\mathcal{B}^{(i)}$ et $\mathcal{B}_\star^{(i)}$ respectivement ($i \in \{I, M, D, R, F\}$, chaque base étant indexée par l'initiale de son nom). On note $\mathcal{S} = \bigcup_i^{\{I, M, D, R, F\}} \mathcal{B}^{(i)}$, c'est-à-dire l'union de toutes nos bases.
- Le modèle entraîné sur $\mathcal{B}^{(i)}$ est indiqué par $\mathcal{M}[\mathcal{B}^{(i)}]$. Ses prédictions sur la base $\mathcal{B}_\star^{(j)}$ sont notées $\mathcal{M}[\mathcal{B}^{(i)}](\mathcal{B}_\star^{(j)})$, ou $\mathcal{M}_{(i)}^{(j)\star}$ en simplifié.
- Un modèle entraîné sur plusieurs bases de données est indiqué par $\mathcal{M}[\bigcup_i \mathcal{B}^{(i)}]$, ou, dans la version simplifiée, par $\mathcal{M}_\mathcal{S}$ quand l'entraînement se fait sur les cinq bases combinées.
- On note $D(\mathcal{M}_{(i)}^{(j)\star}, \mathcal{B}_\star^{(j)})$ le score calculé sur les annotations de $\mathcal{B}_\star^{(j)}$ obtenu par le modèle entraîné sur la base i .

Afin de quantifier la notion de généralisation inter et intrabase, une même architecture est entraînée sur diverses bases de données $\mathcal{B}^{(i)}$ avant d'être évaluée sur $\mathcal{B}_\star^{(j)}$. Cette expérience permet de mettre en évidence la tendance du modèle à reproduire le style d'annotation propre à chaque base de données. En effet, comme cela se constate sur les résultats rapportés sur le tableau 4.4, quelle que soit la paire de bases d'apprentissage/test considérée, le meilleur score (en l'occurrence le *mean Intersection-over-Union* (mIoU)) est toujours obtenu lorsque le jeu d'entraînement et le jeu d'évaluation sont issus de la même base. Autrement dit, formellement, nous observons la relation suivante :

$$D(\mathcal{M}_{(i)}^{(i)\star}, \mathcal{B}_\star^{(i)}) \geq D(\mathcal{M}_{(i)}^{(j)\star}, \mathcal{B}_\star^{(j)}), \forall j \quad (4.2)$$

À première vue, cela pourrait traduire une forme de surapprentissage sur une base de données (et donc d'une mauvaise généralisation). Mais cette interprétation est incomplète. En effet, au sens strict de la définition, nos modèles ne surapprennent pas : nous n'observons pas d'écart

TABLEAU 4.3 Performances comparatives entre les différentes architectures testées. Les modèles sont entraînés et testés sur la base IDRiD (séparée en deux ensembles suivant les règles de la compétition). En bleu sont indiqués nos scores les plus hauts, en gras, ceux de l'état de l'art.

Nos entraînements						
Architecture	Encodeur	MA	HEM	CWS	EX	Moyenne
UNet [188]	ResNet - 18 [189]	0.4892	0.6407	0.7089	0.8512	0.6725
	ResNet - 34	0.4958	0.6457	0.7033	0.8415	0.6716
	ResNest - 50 [190]	0.5041	0.6182	0.6289	0.821	0.6431
	SE ResNet - 50 [52]	0.4730	0.6111	0.6803	0.8233	0.6469
	¹ MIT B2 [191]	0.5123	0.5749	0.7051	0.8408	0.6583
	¹ MIT B4	0.5045	0.6473	0.6959	0.8251	0.6682
UNet++ [192]	ResNet - 18	0.4955	0.6348	0.7063	0.8531	0.6724
	ResNest - 50	0.4900	0.6601	0.6876	0.8019	0.6599
	SE ResNet - 50	0.4906	0.6141	0.7273	0.8169	0.6622
FPN [193]	ResNet - 18	0.4524	0.6476	0.7260	0.8229	0.6622
	ResNest - 50	0.4870	0.6898	0.7529	0.8246	0.6886
	SE ResNet - 50	0.4576	0.6790	0.7396	0.8169	0.6733
	MobileNet V3 [194]	0.3498	0.5828	0.6348	0.7509	0.5796
DeepLab V3+ [195]	ResNet - 18	0.4515	0.6426	0.6967	0.8098	0.6502
	ResNet - 34	0.4238	0.6073	0.6329	0.8329	0.6242
	SE ResNet - 50	0.4623	0.6868	0.7049	0.8204	0.6686
³ Global-Local UNet	ResNest - 50	0.4580	0.6724	0.7080	0.8263	0.6662
État de l'art publié						
L-Seg [136]		0.4630	0.6370	0.7110	0.7950	0.6515
Deep-Bayesian [196]		0.4840	0.5930	0.6410	0.8420	0.6400
² Global-Local UNet [197]		0.5250	0.7030	0.6790	0.8890	0.6990
CARNet [138]		0.5148	0.6389	0.7215	0.8675	0.6857
⁴ Xception-UNet - Collaborative learning [198]		0.4960	0.6936	0.7407	0.8872	0.7044
Classement officiel IDRiD [141]						
Équipe						
VRT		0.4951	0.6804	0.6995	0.7127	0.6469
PATech		0.4740	0.6490	-	0.8850	-
iFLYTEK-MIG		0.5017	0.5588	0.6588	0.8741	0.6483
SOONER		0.4003	0.5395	0.5369	0.7390	0.5539

¹ Encodeur de type ViT suivant le modèle du SegFormer.

² Il s'agit de quatre modèles distincts avec une architecture différente par lésion.

³ Notre réimplémentation est multi-classe (une seule architecture pour les quatre classes).

⁴ Ce modèle combine apprentissage supervisé et faiblement supervisé par entraînement adversarial, utilisant par conséquent beaucoup plus d'images.

flagrant entre les métriques calculées sur $\mathcal{B}^{(i)}$ et $\mathcal{B}_\star^{(i)}$ (c'est-à-dire entre l'ensemble d'entraînement et de test issu de la même base). Plus formellement, $D(\mathcal{M}_{(i)}^\star, \mathcal{B}_\star^{(i)}) \approx D(\mathcal{M}_{(i)}^{(i)}, \mathcal{B}^{(i)})$. Par ailleurs, a minima sur IDRiD, les performances relevées sont de pair avec l'état de l'art, il n'existe donc pas de raison de soupçonner un mauvais protocole d'entraînement. Comment expliquer la détérioration des métriques à travers les différentes bases de test $\mathcal{B}_\star^{(j)}$? Notre hypothèse est que la notion de généralisation n'est a priori pas adaptée pour interpréter ces résultats, car elle repose sur le postulat de données indépendantes et identiquement distribuées selon une distribution $\mathbb{P}(X, Y)$ (images/annotations). À partir du moment où les données appartiennent à des distributions dissociées et bien distinctes, ce postulat ne tient plus et invalide la notion de généralisation. Or notre analyse préliminaire nous le confirme : bien que nous n'y ayons pas explicitement accès, la distribution statistique $\mathcal{B}^{(i)} \sim \mathbb{P}^{(i)}(X, Y)$ semble bien varier selon les bases. Plutôt que de parler de surapprentissage, nous introduisons par conséquent la notion de sursophistication d'un modèle sur une distribution.

Il existe a priori une manière simple de modéliser de manière plus représentative l'ensemble des données : entraîner le modèle à tirer parti de tous les différents styles d'annotations à la fois, c'est-à-dire lui faire apprendre la distribution $\bigcup_i \mathcal{B}^{(i)} \sim \mathbb{P}(X, Y)$. Ce modèle est noté $\mathcal{M}_S$ et son entraînement se fait simplement sur la combinaison de toutes les bases d'apprentissage à notre disposition. Dans nos travaux, celle-ci est extrêmement simple, il s'agit d'une union des ensembles, sans ré-équilibrage. Il existe donc une forte disparité de représentativité de chaque base en termes de nombre d'échantillons ; mais il n'est cependant pas évident qu'il soit nécessaire de corriger celle-ci. Nous avons tenté d'autres approches plus sophistiquées permettant de combiner les données, mais sans obtenir de résultats probants ; celles-ci sont donc présentées dans l'annexe C.3.

Notons qu'à la différence de la littérature sur la généralisation multidomaines, nous ne définissons pas de domaine source ou cible. En effet, ces termes ont un sens lorsqu'on souhaite adapter un modèle d'un domaine vers l'autre (typiquement un généré artificiellement vers un réel). Pour notre application, nous n'avons aucune information a priori sur la pertinence clinique de tel ou tel style d'annotation donc pas d'incitations à privilégier un domaine par rapport aux autres.

Nos résultats indiquent que l'efficacité de cette approche semble validée : sur trois des bases d'évaluation considérées $\mathcal{B}_\star^{(j)}$, ce modèle surpasse tous les autres :

$$\exists k : D(\mathcal{M}_S, \mathcal{B}_\star^{(k)}) > D(\mathcal{M}[\mathcal{B}^{(k)}], \mathcal{B}_\star^{(k)}) \quad (4.3)$$

Pour les deux autres cas de figure, la baisse des performances est relativement faible (-0.0097 pour MESSIDOR, -0.0082 pour DDR). En revanche, la moyenne des scores est très nette-

ment améliorée (+0.125) par rapport au second meilleur modèle. Une analyse plus fine est également proposée par lésions sur les graphiques de la figure 4.4 ; mais elle ne révèle pas une tendance différente : en moyenne $\mathcal{M}_S$ surpasse les autres modèles. Certes, un tel résultat est attendu sur le score moyen, ce modèle ayant été entraîné sur bien plus de données que tous les autres. Mais il est étonnant que l'amélioration ne soit pas qu'en moyenne, mais également sur les scores calculés séparément sur IDRiD, RET-LES et FGADR, ce que nous résumons formellement par l'équation 4.3. Cette équation relève d'un certain paradoxe : en effet, chaque ensemble $\mathcal{B}_\star^{(j)}$ ayant son propre style d'annotation, il semble impossible pour un modèle adoptant un certain style de maximiser les performances sur tous les autres, en particulier quand la différence de styles est aussi importante qu'entre RET-LES et IDRiD. Rappelons que la première base contient en effet des annotations grossières et beaucoup plus larges, tandis que la seconde détaille très précisément le contour des lésions (et celles-ci sont donc bien plus petites dans cette base). Ce paradoxe nous pousse à explorer plus en détail le comportement du modèle $\mathcal{M}_S$ dans la section qui suit.

Analyse qualitative

En complément de l'analyse quantitative, la figure 4.5 propose un échantillon des prédictions réalisées par les différents modèles. Outre les quelques erreurs (sous ou sursegmentation de certaines structures, confusions entre petites hémorragies et microanévrismes) visibles, il est particulièrement frappant d'observer à quel point le modèle entraîné sur une base imite le style d'annotations de celle-ci. Cet effet est particulièrement visible sur le modèle entraîné sur RET-LES. Ces observations qualitatives nous poussent à nous interroger sur la pertinence des métriques de segmentation utilisées, plus exactement sur la pertinence de les comparer à travers différentes bases et différents styles d'annotations. Pour explorer cette question, en annexe C.2, nous proposons une analyse alternative reposant sur la capacité de détection des lésions et non plus sur leur segmentation.

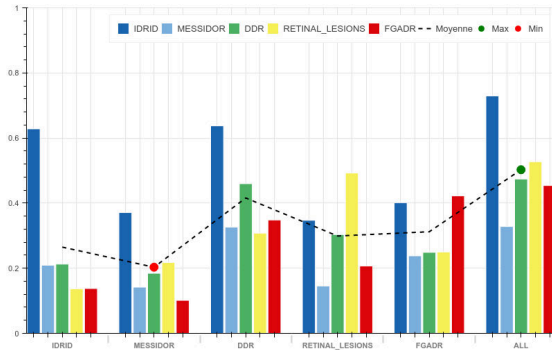
4.3.3 Adaptation de style par Attaque Adversariale

Paradoxe du modèle combiné $\mathcal{M}_S$.

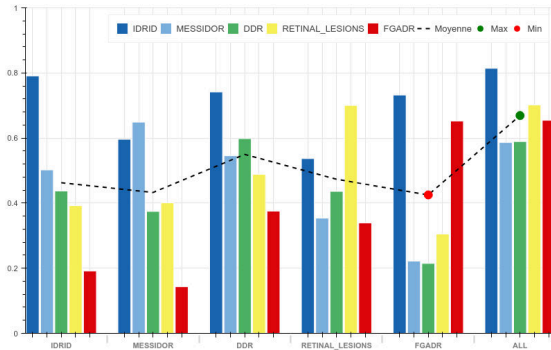
Les résultats quantitatifs et qualitatifs obtenus des modèles $\mathcal{M}[\mathcal{B}^{(i)}]$ sont attendus : chaque modèle adopte le style d'annotation de sa base d'entraînement et obtient naturellement les meilleures performances sur l'ensemble de test correspondant. La surprise vient en revanche du modèle combiné : $\mathcal{M}_S$ parvient à surpasser simultanément trois de ces modèles sur leur propre base, ce qui représente un paradoxe apparent considérant que les styles d'annotations

TABLEAU 4.4 Performance interbases : une même architecture est entraînée sur les différents sous-ensembles des données d’entraînements, dépendamment de leur base d’appartenance.

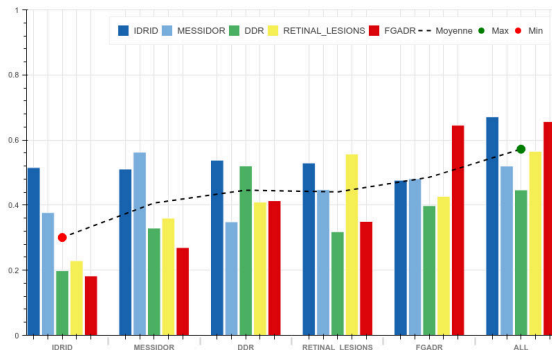
Modèle	Score mIoU sur les ensembles de test					
	IDRiD	MESSIDOR	DDR	RET-LES	FGADR	Moyenne
$\mathcal{M}[\mathcal{B}^{(I)}]$ IDRiD	0.5546	0.3569	0.3192	0.2399	0.2535	0.3448
$\mathcal{M}[\mathcal{B}^{(M)}]$ MESSIDOR	0.3807	0.4570	0.3079	0.2567	0.2157	0.3236
$\mathcal{M}[\mathcal{B}^{(D)}]$ DDR	0.5144	0.3410	0.4290	0.2529	0.2729	0.3620
$\mathcal{M}[\mathcal{B}^{(R)}]$ RET-LES	0.2740	0.2804	0.2660	0.4940	0.2664	0.3162
$\mathcal{M}[\mathcal{B}^{(F)}]$ FGADR	0.3594	0.2616	0.2769	0.2392	0.4591	0.3192
$\mathcal{M}_{\mathcal{S}}$	0.6037	0.4473	0.4208	0.5025	0.4619	0.4872



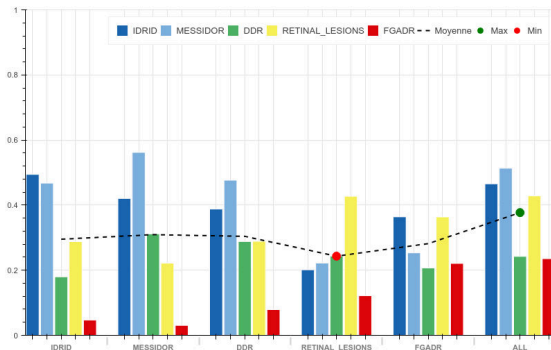
(a) Cotton Wool Spot



(b) Exsudats



(c) Hémorragies



(d) Microanévrismes

FIGURE 4.4 Performance de généralisation par lésion. Suivant l’axe des abscisses sont indiquées les bases d’entraînements, en ordonnées, le score par base d’évaluation. Nous adoptons la métrique AUC sous la courbe Précision/Sensibilité.

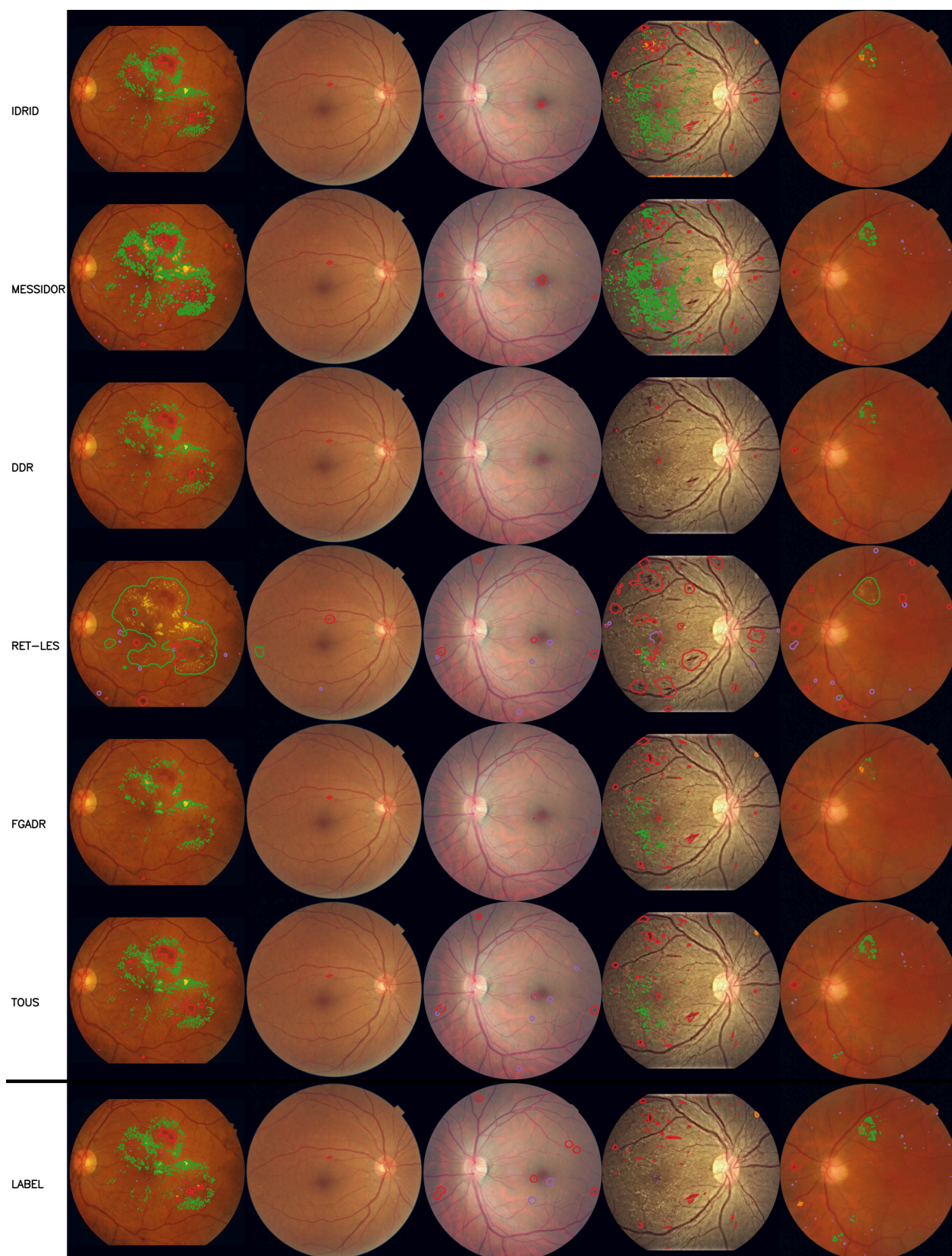
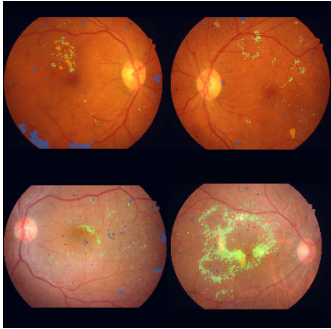
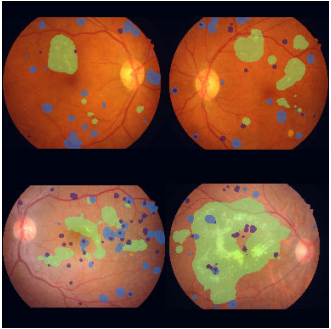
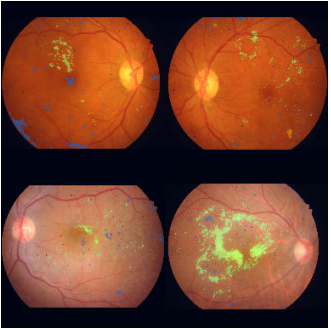
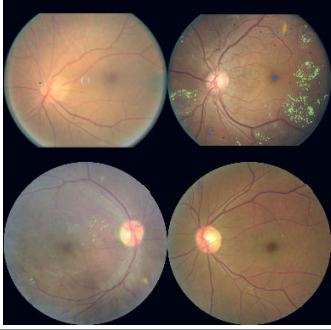
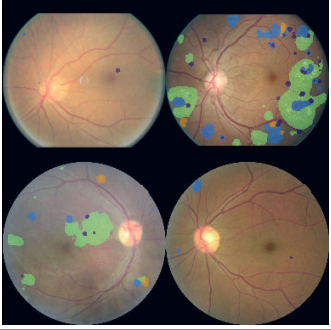
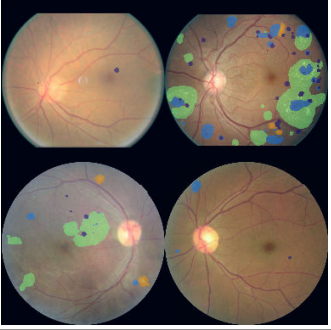


FIGURE 4.5 Résultats qualitatifs de segmentation interbases. Chaque ligne représente un modèle entraîné sur une base. La dernière ligne correspond à la vérité terrain correspondante. Seules les frontières des structures segmentées sont représentées. **CWS**, **EX**, **HEM**, **MA**

TABLEAU 4.5 Illustration du phénomène de la sursophistication d'un modèle sur un style d'annotation. Chaque ligne représente un ensemble de test, chaque colonne un modèle. $\mathcal{B}^{(I)}$ représente la base IDRiD, $\mathcal{B}^{(R)}$ la base RET-LES. La dernière colonne illustre le paradoxe du modèle combiné, capable d'adapter son style à la base de test.

Bases de test	Modèles		
	$\mathcal{M}[\mathcal{B}^{(I)}]$	$\mathcal{M}[\mathcal{B}^{(R)}]$	$\mathcal{M}[\cup_i \mathcal{B}^{(i)}]$
$\mathcal{B}_\star^{(I)}$ (IDRiD)			
$\mathcal{B}_\star^{(R)}$ (RET-LES)			

de chaque base ne sont pas mutuellement compatibles. On s’attendait alors à ce que le modèle adopte le style dominant (ou un style qui représente le meilleur compromis entre chaque base). L’observation des segmentations produites par $\mathcal{M}_S$ résout ce paradoxe (voir le tableau 4.5) non sans soulever d’autres questions : on y voit que le modèle combiné est capable d’ajuster son style de segmentation en fonction de la base d’origine de l’image. Précisons bien qu’il s’agit d’images qui n’ont pas servi à l’entraînement du modèle, donc bien issues de l’ensemble d’évaluation. Le réseau n’a donc a priori pas d’information sur le style d’annotation attendu, mais il arrive pourtant à identifier la base d’origine de l’image, et à adapter sa segmentation à un style concordant. Cette capacité lui permet donc d’obtenir des performances maximales sur les distributions distinctes de chaque base. Elle n’en reste pas moins surprenante : le modèle ne prend pour entrée qu’une image et n’a donc pas accès à son origine de manière explicite. L’unique explication est qu’il existe au sein d’une image un marqueur permettant d’identifier sa base d’appartenance. Le modèle $\mathcal{M}_S$ apprend à reconnaître ce marqueur, à le lire à l’inférence et modifier sa sortie en fonction. Et bien qu’inattendu, cet effet peut s’expliquer : les différentes bases contiennent en leur sein des images dont l’intra-variabilité est plus faible que l’intervariabilité entre bases. En effet, au sein d’une même base, les images sont acquises dans des conditions relativement homogènes : c’est-à-dire même matériel d’acquisition et résolution d’acquisition. Par ailleurs, on peut aussi supposer d’une ethnicité relativement homogène au sein de la population représentée dans chaque base. L’hypothèse d’un marqueur de base identifiable dans chaque image n’est donc pas invraisemblable.

Mais une fois ce comportement de $\mathcal{M}_S$ identifié, plusieurs questions se posent :

- Est-il possible d’identifier le marqueur implicite contenu dans l’image qui trahit son origine ?
- Comment le modèle de segmentation lit-il ce marqueur pour ajuster sa prédiction ?
- Finalement, peut-on contraindre de manière prédictible le style de segmentation du modèle en manipulant ce marqueur de base ?

Précisons que par la suite, nous prendrons souvent IDRiD et RET-LES en exemple, car ces deux bases sont les plus radicalement différentes dans leur style d’annotation respectif.

Robustesse de l’identification de base

De manière préliminaire, nous avons voulu tester la sensibilité de la phase d’identification du marqueur lorsqu’on soumet l’image fournie au modèle à diverses perturbations simples. Pour nos expériences, celles-ci sont de deux types : géométriques ou colorimétriques. Dans le premier cas, on sous-échantillonne l’image (par un facteur aléatoire compris entre 1 et 1/4) avant de la redimensionner à son échelle originale. Cette opération est équivalente à l’application d’un filtre passe-bas et simule une résolution d’acquisition variable. Elle vise à

vérifier que le marqueur d'appartenance à une base n'est pas lié à une certaine composante haute fréquence imperceptible présente dans l'image. La seconde opération modifie de manière aléatoire la teinte, la saturation, la luminosité et le contraste de l'image (dans la limite de 20% d'altérations). Enfin, pour évaluer si la distinction entre base ne provient pas de la compression choisie des images, nous étudions l'effet de celle-ci dans un scénario exacerbé (réduction par compression JPEG à 40% du poids en octet de l'image originale). Comme en témoignent les résultats sur le tableau 4.6, aucune de ces approches ne perturbe le style de segmentation choisi par le réseau de manière probante, à l'exception dans une moindre mesure de la compression JPEG.

Classification de l'origine de l'image

L'approche par perturbations simples ne permettant pas d'expliquer le comportement du modèle, nous adoptons une méthode inspirée par les travaux de Alain et al. [199]. Pour comprendre les caractéristiques lues par un modèle, ces derniers proposent d'inspecter les différentes couches d'un réseau en utilisant des sondes linéaires prenant pour entrée les cartes caractéristiques issues de chaque couche et tentant de les classifier. La linéarité de ces sondes n'est pas en soi indispensable, mais elle confère un caractère interprétable. Par ailleurs, il est pertinent de garder un modèle simple de sonde, car ce n'est pas son pouvoir discriminant qui présente un intérêt, mais celui des caractéristiques issues du réseau inspecté.

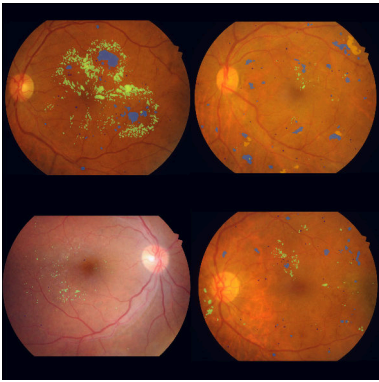
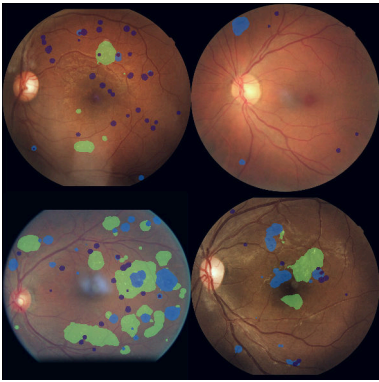
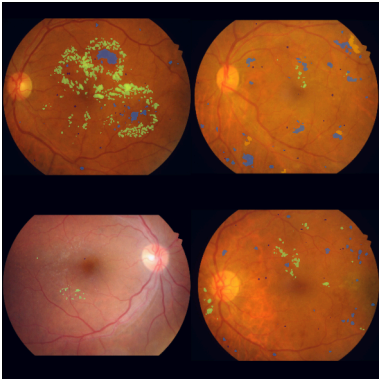
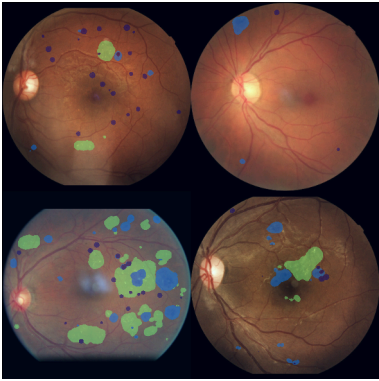
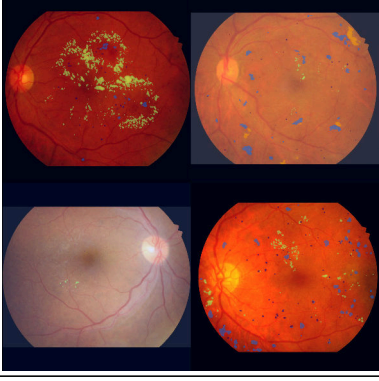
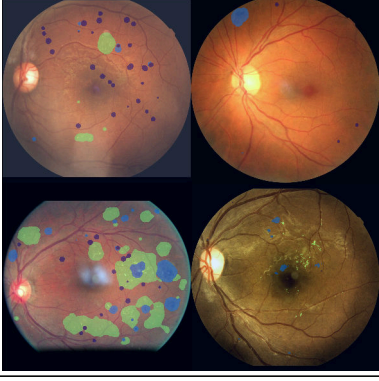
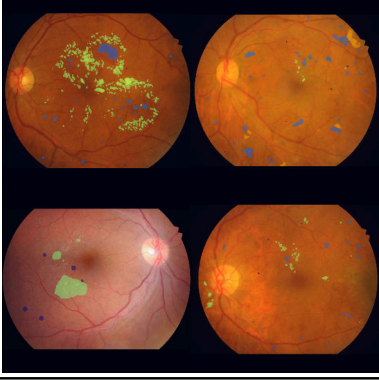
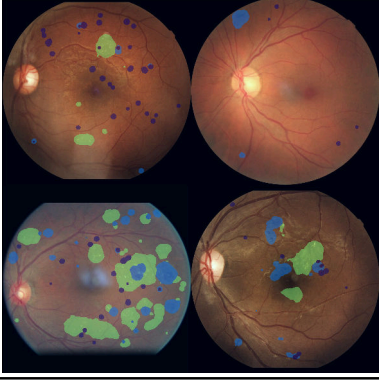
De façon équivalente, nous proposons d'entraîner un micromodèle pour sonder les caractéristiques extraites de l'encodeur du modèle de segmentation. Ce dernier est gelé, autrement dit ses poids ne seront pas modifiés. Le positionnement de la sonde (c'est-à-dire la profondeur de la couche sondée dans l'encodeur) est un paramètre expérimental. La sonde a pour objectif de prédire l'origine de l'image d'entrée, c'est-à-dire sa base d'appartenance parmi les cinq que nous utilisons. Le modèle de sonde employé n'est pas linéaire pour des raisons pratiques de convergence, mais il repose sur une architecture extrêmement simple. Nous l'illustrons sur la figure 4.6. La sonde, dénotée s , est entraînée à minimiser la fonction d'entropie croisée :

$$\mathcal{L}(x_n, y_n) = \sum_{c=1}^C \log s(x_n)_c \cdot y_{n,c} \quad (4.4)$$

où x_n est l'image d'entrée, $s(x_{n,c})$ la probabilité d'appartenance à la base de données c prédit par le modèle et $y_{n,c}$ la vérité terrain.

En pratique, nous avons entraîné séparément différentes sondes, une par couche d'encodeur. Chacune d'entre elles est entraînée sur trois époques sur l'ensemble des données $\bigcup_i \mathcal{B}^{(i)}$. Les résultats sont fournis sur le graphique de la figure 4.7. Pour en faciliter l'interprétation, nous

TABLEAU 4.6 Pour identifier le marqueur de base de données dans l'image, des perturbations géométriques ou colorimétriques sont appliquées à celle-ci. Ces perturbations n'ont pas d'effet sur le style de segmentation choisi par le modèle. Marginalement, le niveau de compression peut avoir un léger effet sur $\mathcal{B}_\star^{(I)}$.

Modèles	Bases	
	$\mathcal{B}_\star^{(I)}$ (IDRID)	$\mathcal{B}_\star^{(R)}$ (RET-LES)
$\mathcal{M}[\cup_i \mathcal{B}^{(i)}]$ (Référence)		
$\mathcal{M}[\cup_i \mathcal{B}^{(i)}]$ + Variation d'échelle (résolution $\times 25\%$)		
$\mathcal{M}[\cup_i \mathcal{B}^{(i)}]$ + Variation colorimétrique (saturation, teinte, contraste, luminosité)		
$\mathcal{M}[\cup_i \mathcal{B}^{(i)}]$ + Compression JPEG (40%)		

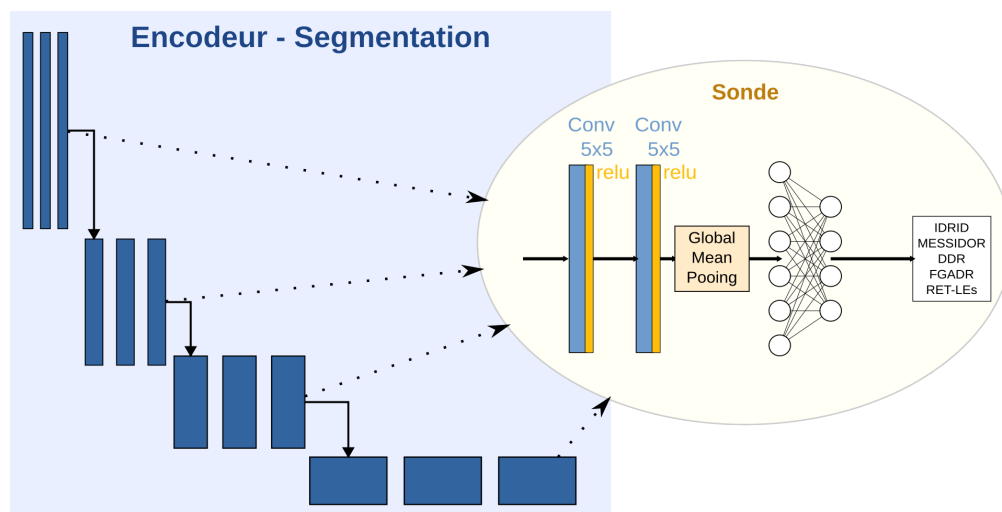


FIGURE 4.6 Modèle de sonde utilisé pour la classification de l'origine des données. Les flèches pointillées indiquent les différents points de connexion, mais il s'agit bien de quatre sondes distinctes, bien que basées sur la même architecture.

avons équilibré les cinq classes (correspondant chacune à une base de données). Avec près de 97% des images correctement classifiées dans le meilleur cas, force est de constater que la classification de l'origine des données ne pose pas problème à la sonde. La performance tend à s'améliorer fortement vers les couches les plus profondes du réseau de segmentation. Cela confirme la capacité du modèle à progressivement décoder ce marqueur dans l'image d'entrée. Dans le détail, notons que sur le graphique de la figure 4.7, l'indice 0 correspond à l'image brute. Rien qu'à partir de celle-ci, la sonde prédit la base d'appartenance d'origine avec une exactitude de près de 60%. Ce chiffre grimpe jusqu'à 97% pour la sonde placée sur l'avant-dernière couche de l'encodeur. Pour celle-ci, nous étoffons notre analyse en calculant la matrice de confusion des prédictions réalisées, rapportée sur la figure 4.8. Dans le détail par classe, on observe que la plus grande source d'erreur provient d'images de MESSIDOR et IDRID (respectivement 28.4% et 17.2% de ces deux bases) prédites comme appartenant à DDR par la sonde. Cela témoigne a priori d'une grande proximité entre ces trois bases, en accord avec notre analyse préliminaire de la section 4.0.2. Quoiqu'il en soit, l'existence d'un identifiant indiquant la base d'origine inscrit dans l'image et décodé progressivement par le modèle de segmentation se confirme à travers nos résultats.

Adaptation par conversion adversariale La facilité de classification de l'origine de l'image par la sonde interroge sur les caractéristiques lues dans celle-ci permettant cette classification. Répondre à cette question offre indirectement une opportunité de premier choix : est-il possible de modifier l'image d'entrée de telle sorte à tromper la sonde ? Si tel est le cas,

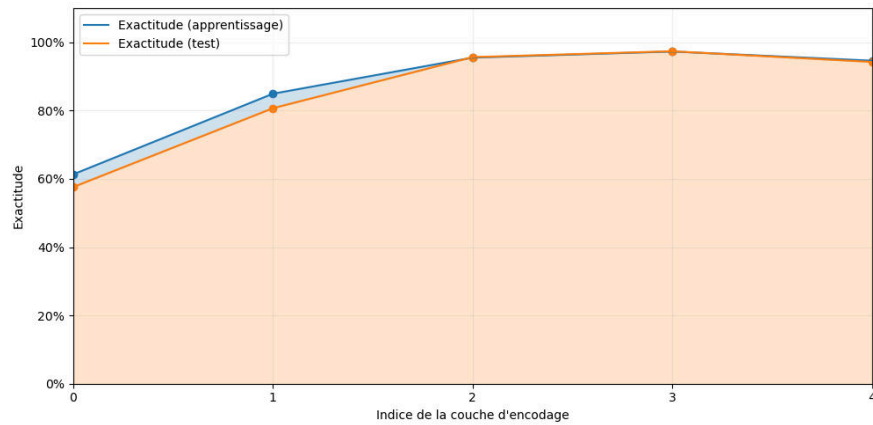


FIGURE 4.7 Performance de classification de l'origine des images ($\mathcal{B}^{(i)}$) suivant la couche d'encodage sondée dans le modèle de segmentation.

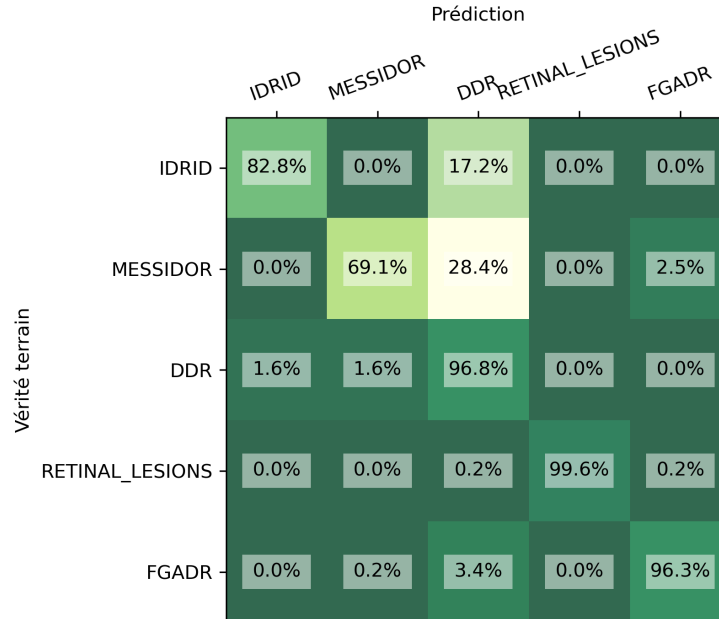


FIGURE 4.8 Matrice de confusion obtenue à partir des prédictions de la sonde située dans l'avant-dernière couche d'encodage du modèle de segmentation.

l'effet se répercute-t-il sur le modèle de segmentation (décodeur inclus) ? Enfin, ce faisant, peut-on le contraindre à adopter un style d'annotations choisi ?

Pour étudier ces questions, nous avons mis au point une expérience basée sur le principe des attaques adversariales. Cette technique vise à modifier imperceptiblement l'entrée d'un réseau de telle sorte à le tromper dans sa classification et à détourner sa prédiction originale vers une classe cible choisie. Dans notre cas, l'attaque est portée sur la sonde de classification dans le but de lui faire prédire une base de données T choisie pour n'importe quelle image qui lui est proposée. Les travaux existants sur les attaques adversariales suggèrent diverses modifications de l'image pour atteindre cet objectif. La plus courante est de rétropropager les gradients de la fonction de coût évalué sur la cible T jusqu'à l'entrée. On modifie ensuite celle-ci par descente de gradients dans la direction qui minimise ce coût (et maximise donc la prédiction T). Il existe une abondante littérature sur le sujet, en général orientée vers la question de la cybersécurité des réseaux de neurones, car le principe des attaques adversariales repose sur l'exploitation d'une de leur vulnérabilité intrinsèque (la sensibilité au gradient). Notre application est transverse à ces questions, car notre objectif est de caractériser la capacité d'adaptation à divers domaines de distribution. Par ailleurs, nous souhaitons bien mettre l'accent sur le fait que l'attaque adversariale a ici pour but de tromper la sonde et non le modèle de segmentation. L'effet sur ce dernier, bien que nous verrons qu'il est radical, en est une conséquence indirecte.

Nous utilisons l'approche de Madry et al. [200] pour concevoir notre attaque : soient x l'image initiale (issue de $\mathcal{B}_\star$) et t la classe que l'on souhaite faire prédire à la sonde. La création d'une image adversariale s'obtient suivant la procédure itérative suivante :

$$\begin{aligned} x_0 &= x \\ x_{k+1} &= \mathcal{S}_\epsilon(x_k - \alpha \cdot \text{signe}(\nabla \mathcal{L}(x_k, t))), k \in \{1, \dots, K\} \end{aligned} \quad (4.5)$$

où $\mathcal{S}_\epsilon$ est une fonction qui restreint la perturbation dans une boule centrée sur x de rayon ϵ . On reconnaîtra ici une équation très similaire à celle de la descente de gradients ; précisons simplement qu'ici, seule l'image d'entrée est modifiée et non les poids ni de la sonde ni du réseau de segmentation. L'équation 4.5 se répète K fois et l'image résultante x_K est théoriquement censée perturber la prédiction de la sonde vers la classe choisie. Par ailleurs, en contraignant un ϵ petit, on force cette modification à rester imperceptible à l'œil humain. Pour nos expériences, nous utilisons $\epsilon = 0.025$, un pas de modification $\alpha = 0.002$ et $K = 15$ étapes. Du fait du petit ϵ , les perturbations sont a priori imperceptibles à l'œil humain : les valeurs de pixels de l'image étant comprises dans l'intervalle $[-1, 1]$, elles ne peuvent être

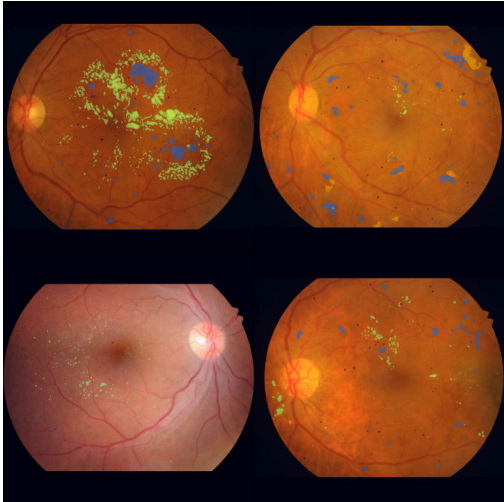
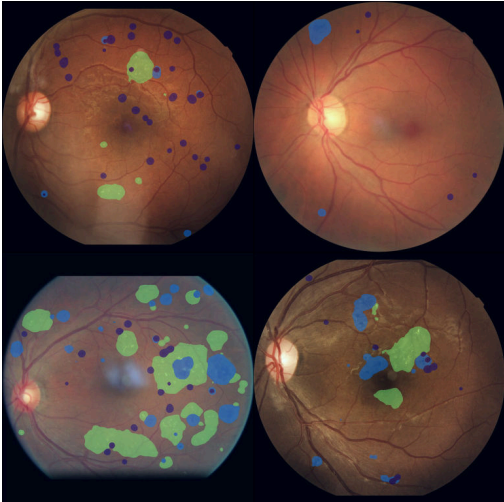
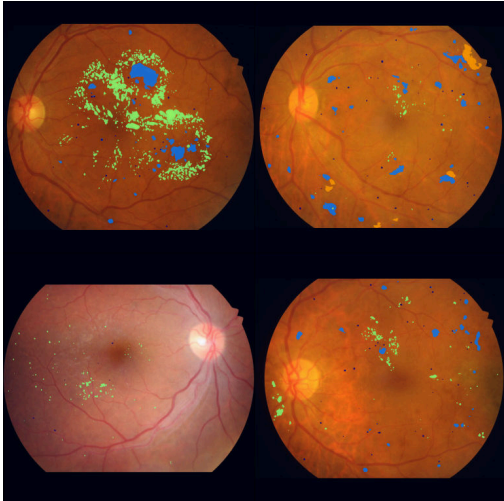
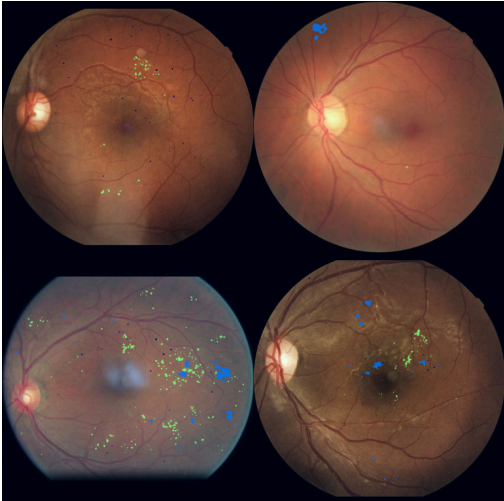
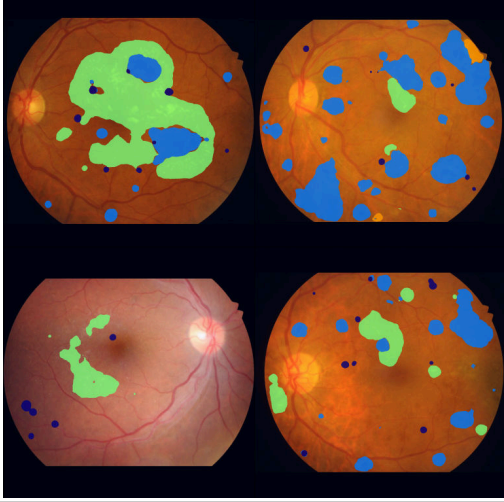
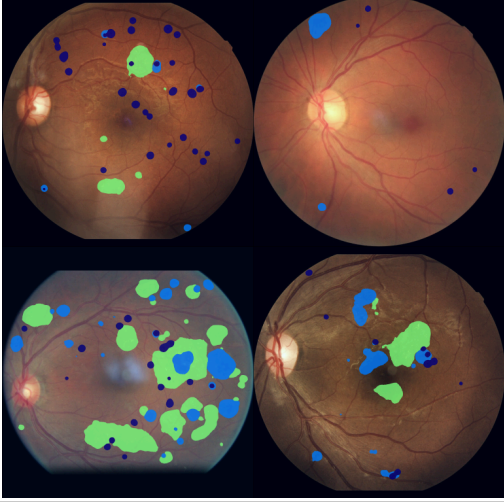
modifiées au maximum que de 0.125%.

Pour simplifier la lecture des résultats, nous introduisons une nouvelle notation : $\mathcal{B}^{(i)} \Rightarrow \mathcal{B}^{(j)}$ indique une conversion adversariale des images de la base source $\mathcal{B}^{(i)}$ vers la base cible $\mathcal{B}^{(j)}$. Ainsi $M(\mathcal{B}_\star^{(j)} \Rightarrow \mathcal{B}^{(i)})$ signifie que le modèle M est testé sur des images appartenant à $\mathcal{B}_\star^{(j)}$ qui ont été modifiées pour correspondre à la base $\mathcal{B}^{(i)}$. Nous pouvons quantifier le succès de conversion, en observant la proportion d'images classées T par la sonde après conversion $\bigcup_i \mathcal{B}_\star^{(i)} \Rightarrow \mathcal{B}^{(T)}$. Les résultats sont rapportés sur la figure 4.9. Dépendamment de la cible T , le taux de conversion réussie varie entre 82.5% (vers IDRID) et 65.2% (vers MESSIDOR). Dans le détail, les images issues FGADR sont celles les plus difficilement converties, avec 15.3% de conversion ratée provenant d'images de cette base (pour la conversion vers IDRID), 21.4% (vers MESSIDOR) et 33.5% (vers RETINAL-LESIONS). Si on exclut FGADR de nos images de test, le taux de succès de conversion moyen (toutes bases confondues) est de 90.5%, le taux le plus faible étant pour la conversion vers FGADR (79% de succès) et le plus élevé pour RETINAL-LESIONS (98.8%).

Cependant, tromper la sonde n'est pas notre objectif en soi. Dans le contexte de notre étude d'intérogénéralisation, il est plus intéressant d'observer comment l'image transformée impacte la prédiction du modèle de segmentation. Dans le tableau 4.7, nous illustrons l'effet de la conversion sur le modèle. La sonde choisie est connectée à l'avant-dernière couche de l'encodeur et nous avons expérimenté avec deux types de conversions (vers IDRID et vers RETINAL-LESIONS). Considérant que toutes les segmentations du tableau sont réalisées avec le même modèle inchangé et sur des images de test, les résultats obtenus se révèlent déroutants : les modifications adversariales apportées à l'image d'entrée permettent de profondément modifier le style de segmentation. Ce faisant, nous obtenons une très grande flexibilité sur le style de segmentation en sortie : un même modèle peut-être configuré, sans avoir à le modifier et donc sans réentraînement coûteux, à adopter un style de segmentation parmi plusieurs alternatives. Qui plus est, le processus de conversion est stable : la conversion $\mathcal{B}^{(i)} \Rightarrow \mathcal{B}^{(i)}$ ne modifie pas la segmentation du modèle.

Validation quantitative L'efficacité du processus de conversion des images reste encore à démontrer. À défaut d'avoir plusieurs styles d'annotations pour une même image, nous disposons de modèles dont nous savons qu'ils ont chacun adopté un style d'annotation spécifique (les $\mathcal{M}[\mathcal{B}^{(i)}]$). Ceux-ci nous serviront pour notre protocole d'évaluation. Notre objectif est d'éprouver la fiabilité du processus de conversion adversariale dans un contexte pseudo-clinique, donc sur des images qui n'appartiennent à aucune des bases utilisées jusqu'à présent. Nous prélevons pour cela un sous-ensemble de 100 images issues de la base APTOS, pour lesquelles nous ne disposons pas d'annotations de lésions. On note $\mathcal{K}$ cet ensemble. À par-

TABLEAU 4.7 L'attaque adversariale de la sonde de classification modifie de manière imperceptible à l'œil nu l'image d'entrée. En revanche, elle permet de drastiquement modifier la segmentation pour adopter un style au gré de l'utilisateur.

Modèles	Bases			
	$\mathcal{B}_\star^{(I)}$		$\mathcal{B}_\star^{(R)}$	
$\mathcal{M}[\cup_i \mathcal{B}^{(i)}]$ (Réf- erence)				
$\mathcal{M} \Rightarrow \mathcal{B}^{(I)}$				
$\mathcal{M} \Rightarrow \mathcal{B}^{(R)}$				

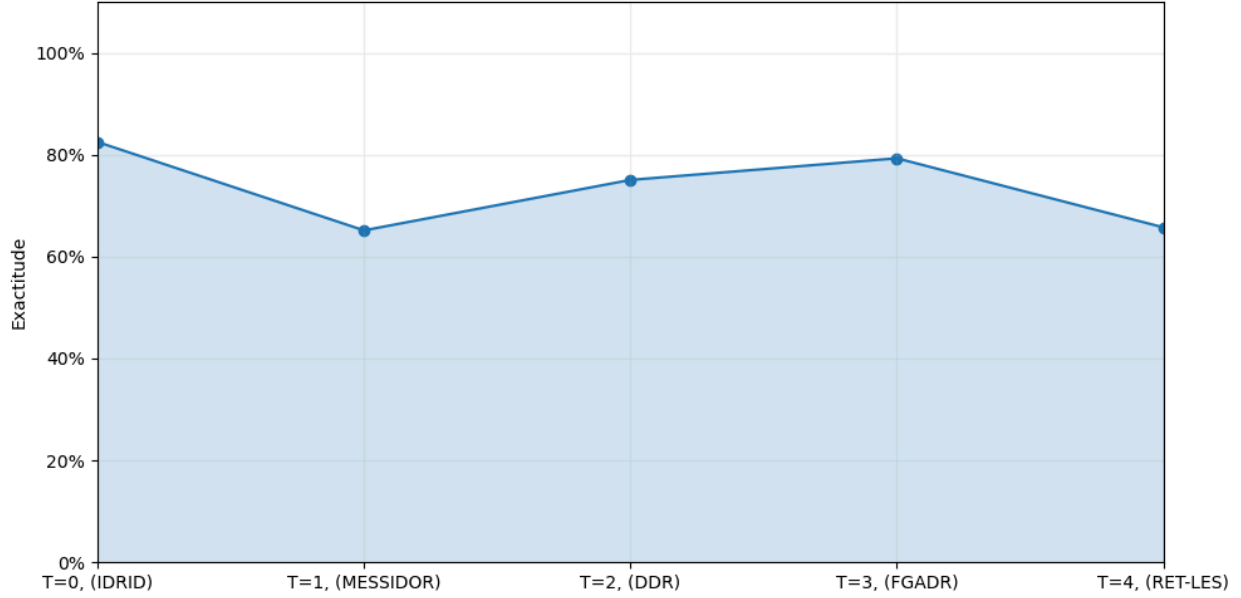


FIGURE 4.9 Pourcentage d’images des bases $\bigcup_i \mathcal{B}^{(i)}$ classées T par la sonde après conversion $\bigcup_i \mathcal{B}^{(i)} \Rightarrow \mathcal{B}^{(T)}$. Cette proportion correspond au taux d’images ayant trompé la sonde après conversion vers la cible T visée.

tir de chacun des modèles $\mathcal{M}[\mathcal{B}^{(i)}]$, on génère cinq styles d’annotation de segmentations qui nous serviront de référence. Elles seront comparées avec les segmentations de $\mathcal{M}[\bigcup_i \mathcal{B}^{(i)}]$ après conversion adversariale vers chacune des bases. Formellement, notre validation repose donc sur l’évaluation de la métrique D entre les modèles spécialisés et le modèle de référence après conversion des images vers le domaine de chaque modèle spécialisé :

$$D(\mathcal{M}_{(j)}(\mathcal{K}), \mathcal{M}_S(\mathcal{K} \Rightarrow \mathcal{B}^{(t)})), \forall j, t \quad (4.6)$$

À l’instar de nos expériences précédentes, nous utilisons la mIoU comme métrique de comparaison. Les résultats sont synthétisés sur la matrice représentée sur la figure 4.10. L’intérêt de ces résultats apparaît par une lecture ligne à ligne de la matrice fournie. En particulier, on observe que ;

$$D(\mathcal{M}_{(j)}(\mathcal{K}), \mathcal{M}_S(\mathcal{K} \Rightarrow \mathcal{B}^{(j)})) \geq D(\mathcal{M}_{(j)}(\mathcal{K}), \mathcal{M}_S(\mathcal{K} \Rightarrow \mathcal{B}^{(t)})), \forall j, t \quad (4.7)$$

Autrement dit, après conversion vers la classe T , le modèle $\mathcal{M}_S \Rightarrow T$ s’approche systématiquement le plus du modèle $\mathcal{M}_T$. Ce résultat confirme l’efficacité de l’approche adversariale pour modifier au choix le style de segmentation du modèle.

	$\mathcal{K} \Rightarrow \mathcal{B}^{(I)}$	$\mathcal{K} \Rightarrow \mathcal{B}^{(M)}$	$\mathcal{M}[\bigcup_i \mathcal{B}^{(i)}](\mathcal{K} \Rightarrow \mathcal{B}^{(j)})$ $\mathcal{K} \Rightarrow \mathcal{B}^{(D)}$	$\mathcal{K} \Rightarrow \mathcal{B}^{(R)}$	$\mathcal{K} \Rightarrow \mathcal{B}^{(F)}$
$\mathcal{M}[\mathcal{B}^{(I)}](\mathcal{K})$	88.2	86.8	88.4	73.6	81.4
$\mathcal{M}[\mathcal{B}^{(M)}](\mathcal{K})$	64.9	76.4	67.6	58.1	63.3
$\mathcal{M}[\mathcal{B}^{(D)}](\mathcal{K})$	89.2	84.2	89.6	78.6	87.5
$\mathcal{M}[\mathcal{B}^{(R)}](\mathcal{K})$	77.4	72.1	78.8	87.4	85.0
$\mathcal{M}[\mathcal{B}^{(F)}](\mathcal{K})$	75.6	68.6	73.2	68.8	75.1

FIGURE 4.10 Cette matrice représente le score mIoU obtenu en comparant les prédictions réalisées par chacun des modèles spécialisés $\mathcal{M}_{(j)}$ et celles réalisées par le modèle $\mathcal{M}_S$ après conversion adversariale des images d'entrées vers chacune des bases.

4.4 Gradation automatique de la rétinopathie diabétique assistée par segmentation

La dernière étape méthodologique de ce travail prend du recul sur la question de la qualité de la segmentation et interroge la pertinence de celle-ci dans un contexte clinique. Elle vise à répondre de manière plus pragmatique sur la qualité des segmentations produites par les différents modèles, non pas sous le prisme des performances pixel par pixel, mais sous l'angle de la pertinence de celles-ci pour le diagnostic de l'image globale. Nous avons fait pour cela le choix d'une approche par graphe pour classifier une image de fond d'œil en différents grades de maladie (la rétinopathie diabétique) à partir de ses lésions. Une telle approche est une contribution originale sans équivalents dans la littérature à notre connaissance. Pour la justifier, la section suivante 4.4.1 est une parenthèse dans la description de notre méthodologie et présente les expériences préliminaires qui ont motivé l'approche par graphe.

4.4.1 Expériences préliminaires : le problème de la classification assistée par segmentation

Il n'existe pas de façons univoques de labelliser les lésions dans une image, ce qu'atteste la variabilité des annotations selon les bases. Dans ce contexte, comment évaluer la qualité d'un modèle entraîné sur ces bases ? Nous proposons de répondre à cette question par procuration en reformulant le problème : étant donnée une carte de segmentation générée par un modèle, peut-on retrouver le diagnostic de l'image associée ? Corollairement, quel style d'annotation est le plus efficace pour aboutir à un diagnostic ?

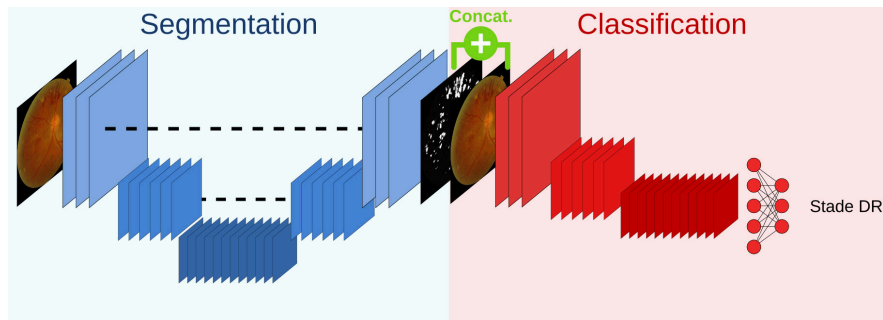
Surprenamment, cette question est rarement posée dans la littérature. Certains travaux combinent l'étape de segmentation à celle de classification [140, 201, 202] et démontrent généralement des performances améliorées par l'ajout de cette combinaison. Mais dans le détail, il s'agit en général de gains relativement marginaux. Il est par ailleurs établi que les réseaux de neurones sont particulièrement doués pour classifier des images sans présegmentation. Ainsi, l'ajout des marqueurs segmentés peut être simplement vu comme une forme de guidage de l'apprentissage, comme une supervision renforcée par la distillation de l'expertise du modèle de segmentation. Mais aucun de ces travaux n'étudie l'impact de la qualité des segmentations sur la classification, autrement dit, il n'y est pas du tout évident qu'une mauvaise segmentation entraînera une mauvaise classification.

Pour y voir plus clair, nous avons mené des expériences préliminaires, en combinant simplement un réseau de classification et de segmentation suivant différentes approches, librement inspirées par la littérature. Trois combinaisons sont expérimentées, représentées sur la figure 4.11 :

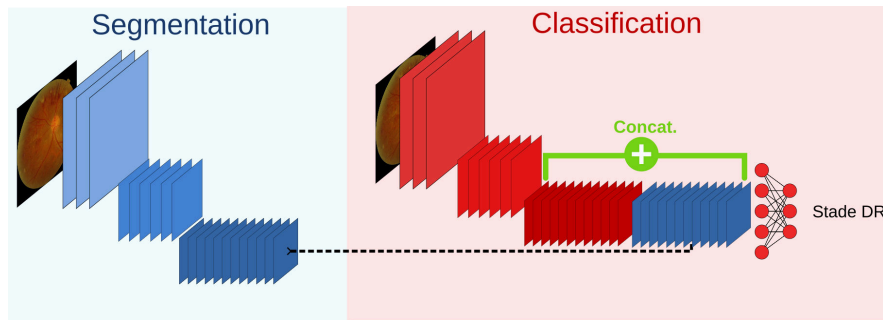
1. Combinaison séquentielle : la sortie du réseau de segmentation est concaténée à l'image originale. Les cartes de segmentations (non binarisées, c'est-à-dire donc les probabilités associées à chaque classe) sont fournies directement à un réseau de classification après concaténation avec l'image d'origine (ou sans dans la version alternative du modèle).
2. Combinaison parallèle : les réseaux de segmentation et de classification reçoivent chacun la même image. Les caractéristiques issues de l'avant-dernière couche de chaque encodeur sont concaténées.
3. Combinaison hybride : ce scénario consiste en un mixte des deux scénarios précédents.

Précision pratique : l'architecture de classification fonctionne sur une résolution d'image de 512×512 (contre 1536×1536 pour la segmentation). Pour pallier cette différence, les cartes issues du réseau de segmentation sont interpolées à la résolution adéquate avant concaténation. Le modèle de segmentation est entraîné séparément, suivant les modalités des sections précédentes. Nous admettons donc disposer de différents modèles aux styles de segmentation distincts. En entraînant la classification sur des jeux de données dédiés, nous avons mesuré la performance du modèle en utilisant le coefficient quadratique κ de Cohen. Les bases de données utilisées sont EyePACS [203] (divisée en ensembles d'apprentissage et de test) et AP-TOS [204] (test uniquement). Ces données, que nous réemployons dans le chapitre suivant, y seront décrites avec plus de détails (en particulier dans la section 5.1.4). Précisons simplement que ces bases, bien plus conséquentes que celles dédiées à la segmentation, fournissent des images et un grade de rétinopathie diabétique suivant une échelle allant de 0 à 4. Les résultats obtenus avec nos différentes combinaisons sont présentés dans le tableau 4.9. À l'inverse de plusieurs tendances de littérature, il y apparaît que l'ajout d'une segmentation n'améliore pas systématiquement les performances de classification. On obtient cependant des combinaisons pour lesquelles celles-ci sont rehaussées par l'intégration du modèle de segmentation. Il est donc possible que seules certaines symbioses classification-segmentation soient pertinentes. Mais cela met en lumière que le bénéfice de la segmentation pour un modèle déjà capable de faire de la classification n'a rien d'une évidence. En particulier, nous avons expérimenté des variantes de nos combinaisons pour lesquelles l'image d'entrée n'est pas fournie au modèle de classification (seules les structures ou caractéristiques du modèle de segmentation sont utilisées). Dans ce scénario, la performance de classification est fortement dégradée.

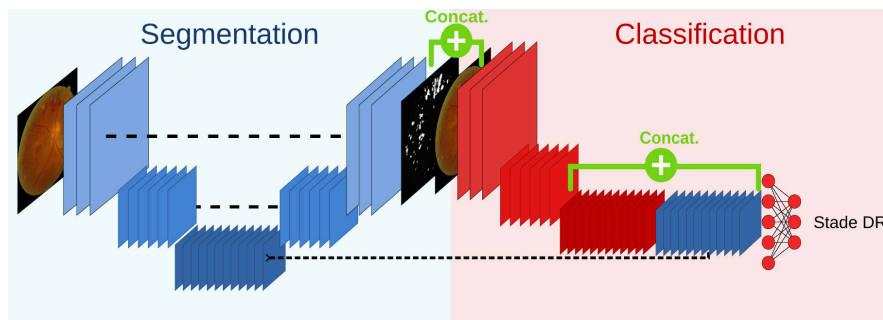
Un second problème restreint les conclusions que l'on peut tirer de modèles CNNs combinant segmentation et classification. Ce problème, indépendant du mode de combinaison choisi, est l'invariance de la classification à la qualité de la segmentation. En réalité, l'entraînement achevé, on peut substituer le modèle de segmentation par un autre au comportement très distinct et n'observer quasiment aucune variation dans les performances de classification à l'inférence. Cela remet fortement en question l'utilité d'une telle combinaison lors de



(a) Réseaux de segmentation et classification mis en série. La sortie de la segmentation est envoyée comme entrée (concaténée à l'image) du réseau de classification



(b) Réseaux de segmentation et classification mis en parallèle, avec fusion des caractéristiques intermédiaires des deux modèles



(c) Combinaison parallèle/série, c'est-à-dire un mixte des approches 4.11a et 4.11b.

FIGURE 4.11 Dans le cadre d'une étude préliminaire sur la combinaison des informations de segmentation et de classification. Trois approches ont été expérimentées, basées sur différentes stratégies de combinaisons des deux réseaux

l'inférence. Pour illustrer notre propos, nous avons donc repris les architectures entraînées précédemment, en changeant le modèle de segmentation à l'inférence, voire en le retirant entièrement. Ces résultats sont donnés dans le tableau 4.8. Ils indiquent que si la segmentation peut aider l'apprentissage, son effet sur l'inférence est très minime.

Ce type de combinaisons entre segmentation et classification ne permet pas de trancher franchement sur la qualité de la segmentation, problématique qui nous intéresse ici.

Compte tenu de ces résultats, nous avons élaboré une méthodologie permettant de s'abstraire complètement de l'image (et des modèles de classification qui l'utilisent). L'objectif est alors de concevoir un modèle qui n'utilise que les lésions segmentées pour prédire son diagnostic. On le contraint ainsi à apprendre presque explicitement les heuristiques de diagnostic préconisées par les autorités de santé (par exemple [205]) et à ignorer d'autres caractéristiques qu'un CNN serait susceptible d'observer dans une image.

4.4.2 Conception et classification d'un graphe rétinien

L'obtention d'une classification de fond d'œil qui ne repose que sur les lésions est une piste originale ; mais son objectif est avant tout d'évaluer la qualité de la segmentation.

La représentation d'une image comme la somme de ses lésions peut se faire de différentes façons, la plus simple étant probablement de définir un vecteur de caractéristiques par lésion et de les concaténer en une seule matrice représentant l'image. Mais il apparaît immédiatement qu'une telle approche ignore la représentation spatiale de la distribution des lésions dans l'image et en particulier leurs relations de voisinage avec certaines structures vitales.

Notre approche consiste à construire un graphe rétinien à partir des lésions segmentées, qui sera ensuite classifié par un nouveau modèle dédié. Pour construire ces graphes, on représente chaque structure segmentée de l'image (c'est-à-dire chaque composante connectée) sous la forme d'un nœud, qui encapsule un certain nombre de caractéristiques descriptives.

L'intérêt d'un tel graphe est multiple : en construisant une connectivité locale entre nœuds

TABLEAU 4.8 Illustration de l'invariance des performances de classification suivant le modèle de segmentation utilisé. La colonne $\emptyset$ correspond au cas où les caractéristiques de segmentation sont mises à zéro.

Approche	Score	$\emptyset$	Modèles de segmentation					
			$\mathcal{M}_I$	$\mathcal{M}_M$	$\mathcal{M}_D$	$\mathcal{M}_R$	$\mathcal{M}_F$	$\mathcal{M}_S$
Parallèle	$\frac{\kappa^{Aptos} + \kappa^{EyePacs}}{2}$	0.843	0.842	0.842	0.842	0.844	0.843	0.842
Série	$\frac{\kappa^{Aptos} + \kappa^{EyePacs}}{2}$	0.804	0.830	0.788	0.833	0.838	0.824	0.836

TABLEAU 4.9 Performances de gradation de la RD obtenues en testant différentes combinaisons du modèle de segmentation et de classification

	Modèles	κ - Aptos	κ - EyePACS	κ Moyen
Sans segmentation	Référence	0.852	0.821	0.837
Classification avec image + segmentation	Réseaux en série	0.867	0.808	0.838
	Réseaux en parallèle	0.854	0.827	0.841
	Réseaux en parallèle et série	0.835	0.810	0.823
Classification sans image	Réseaux en série	0.755	0.705	0.730
	Réseaux en parallèle et série	0.820	0.777	0.796

voisins, il maintient la notion de distribution spatiale des lésions, chaque nœud peut encoder autant d'informations qu'on le souhaite et la représentation de l'image reste in fine extrêmement allégée (car une image contient au maximum quelques centaines de lésions/nœuds, à comparer aux millions de pixels qui la compose). Le modèle qui sera chargé de classifier l'image ainsi représentée a un nombre très restreint de points à manipuler. Lors de la phase d'entraînement d'un tel modèle, celle-ci ne prenait que quelques minutes sur des bases de plusieurs dizaines de milliers d'images comme EyePACS, contre plusieurs heures pour les architectures CNN de la section précédente. De plus, le modèle entraîné est également particulièrement léger. Bien que cela ne soit pas dans nos objectifs, cela favorise indéniablement son déploiement sur des systèmes embarqués.

L'idée du graphe étant posée (en l'état, il s'agit plus exactement d'un nuage de points composé des lésions segmentées), il reste à conceptualiser sa classification. Pour cela, nous proposons deux architectures reposant sur les modèles de type Graph Neural Network (GNN). La première, qui nous sert de référence, est une retranscription quasiment à l'identique de l'architecture pionnière de Qi et al. [206] appelée PointNet++. La seconde est une contribution originale combinant divers opérateurs innovants récemment introduits dans la littérature.

PointNet++

À l'origine, ce modèle est conçu pour la classification et la segmentation de nuages de points 3D, où chaque nœud n'encapsule que ses coordonnées spatiales (x, y, z). Pour notre application, on agrandit le vecteur descripteur de chaque nœud en concaténant :

- Les caractéristiques issues de l'antépénultième couche de l'encodeur du réseau de segmentation extraites à la position spatiale du nœud. Il s'agit ainsi d'un vecteur descripteur de la lésion $\omega \in \mathbb{R}^{512}$
- La taille de la lésion considérée (en pixels) $t \in \mathbb{N}$

— La classe prédite pour la lésion $c \in \{0, 1, 2, 3, 4\}$.

À noter que nous représentons également le fond de la rétine sous la forme d'un unique nœud central. Cela permet de représenter toutes les images, y compris celles n'ayant aucune lésion segmentée. Le vecteur caractéristique h_i du nœud i s'exprime donc comme :

$$h_i = [\omega_i, t_i, c_i] \in \mathbb{R}^{514} \quad (4.8)$$

Par ailleurs, on note $p_i = (x_i, y_i)$ la position du nœud, c'est-à-dire les coordonnées normalisées du centroïde de la lésion correspondante.

L'architecture PointNet++ est organisée en blocs qui réalisent les trois mêmes étapes :

1. La construction d'un graphe dynamique à partir du nuage de points des lésions. Pour chaque instance, une connectivité est générée de telle sorte que chaque nœud soit connecté à ses k plus proches voisins (au sens de proximité spatiale).
2. Les représentations des nœuds adjacents sont agrégées entre elles. Il s'agit de l'étape de *transmission de message* propre aux architectures de type GNN. Cette étape s'interprète comme une couche convolutive d'un CNN. Tout comme pour les architectures CNNs, on organise plusieurs couches au sein d'un bloc de même résolution.
3. Similairement au *pooling* dans les CNNs, un ré-échantillonnage du graphe est appliqué, suivant l'algorithme du Point le Plus Éloigné (*Farthest Point Sampling*)

L'agrégation des caractéristiques entre nœuds connectés utilise un simple réseau de neurones à deux couches. Soient $\mathcal{N}(i)$ l'ensemble des nœuds connectés à i . La transmission de message d'une couche l à la suivante se fait selon l'équation :

$$h_i^{(l+1)} = \max_{j \in \mathcal{N}(i)} \text{MLP}(\text{ReLU}(\text{MLP}([h_i^{(l)}, p_j - p_i]))) \quad (4.9)$$

En fin de réseau, toutes les cartes de caractéristiques sont agrégées en une seule (*Global Pooling*) qui est transmise à une couche de classification linéaire. Les différentes étapes intervenant dans l'architecture du réseau sont illustrées sur la figure 4.12.

Le PointNet++ attache donc une importance particulière à la position relative des lésions dans l'image, à la fois pour la construction initiale du graphe d'adjacence, le calcul de la transmission des messages (qui explicite la distance relative entre nœuds dans les caractéristiques du message transmis) et le ré-échantillonnage entre blocs.

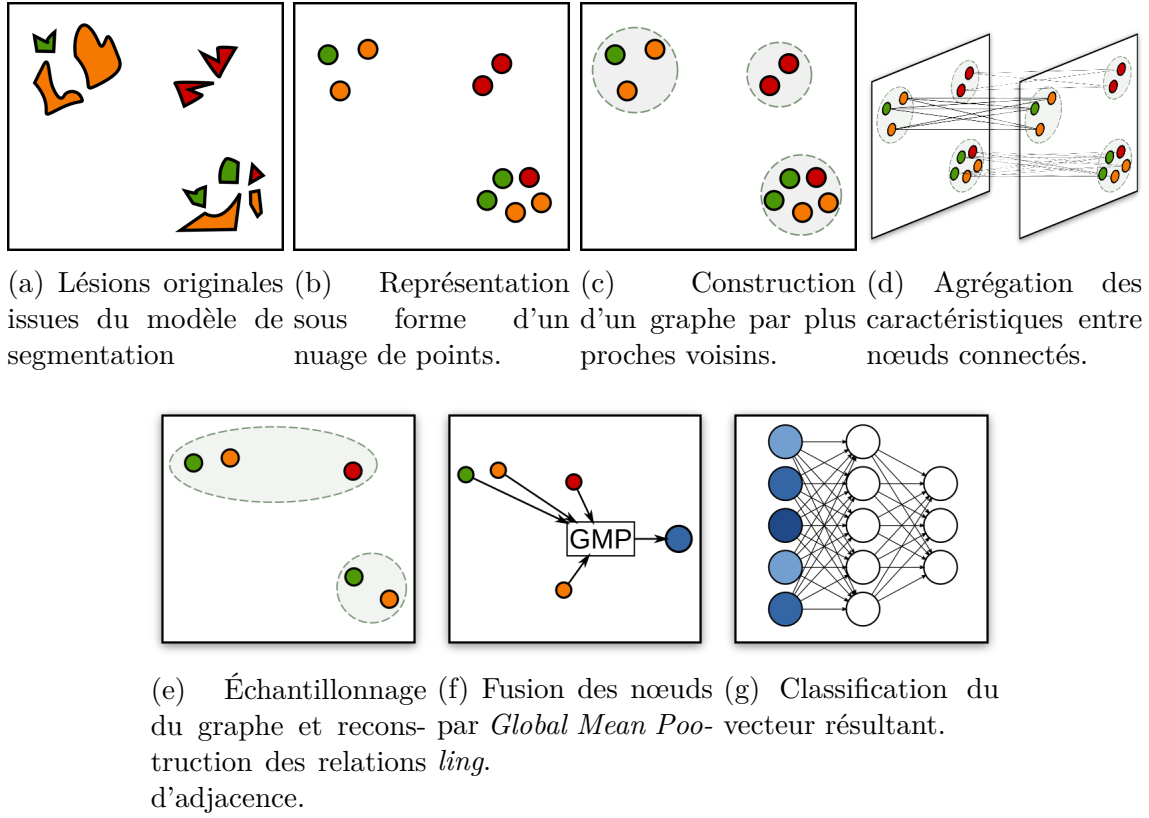


FIGURE 4.12 Les différentes étapes mises en jeu dans l'architecture PointNet++, adaptée à la classification du graphe de lésions. Les étapes de 4.12c à 4.12e représentent les opérations au cœur d'une couche du réseau, répété K fois (selon l'architecture)

Attention-GNN

La seconde architecture que nous proposons est similaire à celle du PointNet++. Structurellement, elle repose sur le même principe de blocs au sein desquels sont transmis des messages de nœuds en nœuds. En revanche, nous remplaçons l'opérateur de transmission des messages du PointNet++ par le GATv2 (pour *Graph Attention Network*) introduit par Brody et al. [207]. Cet opérateur utilise un mécanisme d'attention similaire à celui des Transformers [53] pour gérer la transmission de message entre nœuds adjacents, à la seule différence que l'attention calculée est dynamique. Dans le détail, pour chaque nœud i , l'équation de mise à jour de son vecteur de représentation dans la couche l est donnée par :

$$h_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \mathbf{W}h_j\right) \quad (4.10)$$

où $h_i^{(l)} \in \mathbb{R}^d$. Le coefficient d'attention α_{ij} indique l'importance du nœud j vis-à-vis du nœud i . Pour le calculer, on construit en premier lieu un score par arête unissant deux nœuds adjacents.

$$e(\mathbf{h}_i, \mathbf{h}_j) = \mathbf{a}^T \text{ReLU}(\mathbf{W} \cdot [\mathbf{h}_i || \mathbf{h}_j]) \quad (4.11)$$

où $\mathbf{a} \in \mathbb{R}^{2d'}$, $\mathbf{W} \in \mathbb{R}^{2d' \times 2d}$ sont des paramètres appris du modèle et $||$ représente l'opérateur de concaténation. L'obtention d' α_{ij} se fait par une normalisation de type softmax de ce score d'arêtes :

$$\alpha_{ij} = \frac{\exp(e(\mathbf{h}_i, \mathbf{h}_j))}{\sum_{j' \in \mathcal{N}_i} \exp(e(\mathbf{h}_i, \mathbf{h}_{j'}))} \quad (4.12)$$

Cette architecture est donc plus complexe que le PointNet++, mais structurellement, elle respecte les mêmes principes.

4.4.3 Classification de la rétinopathie diabétique par GNN

Afin de pousser l'exploration sur la qualité des lésions, nous proposons de reconstituer le diagnostic de l'image à partir des cartes de segmentation produites par les différents modèles $\mathcal{M}[\mathcal{B}^{(i)}]$. Deux architectures de classification sont comparées, telles que décrites dans la section méthodologique. L'entraînement est réalisé sur les 35126 images de l'ensemble d'apprentissage de la base EyePACS. 5% de ces images sont utilisées pour la validation du modèle. Le modèle qui obtient le meilleur score de validation au cours de 250 époques est ensuite évalué sur les bases Aptos et EyePacs (test). Les résultats sont reportés dans le tableau 4.10. En moyenne, quel que soit le modèle de segmentation (hormis celui entraîné sur IDRiD), l'architecture Att-GNN obtient les meilleures performances. Notamment, ce constat est particulièrement notable sur l'ensemble de test EyePacs, composé d'images de bien moins bonne qualité que sur Aptos.

Dans le détail, on observe également la supériorité (en moyenne sur les deux ensembles) des performances obtenues avec $\mathcal{M}_S$ obtenant respectivement 0.735 (Att-GNN) et 0.682 (PointNet++), suivi de $\mathcal{M}[\mathcal{B}^{(R)}]$ (scores respectivement de 0.714 et 0.663).

Relativement à ce que produit un CNN à partir d'une image, ces scores ne sont pas élevés, notamment par rapport à ceux obtenus dans l'étude préliminaire de la section 4.4.1. En les analysant plus finement, il apparaît que les modèles par graphe peinent surtout à distinguer certaines classes, en particulier la rétinopathie proliférative. Ce phénomène n'est pas surprenant, dans la mesure où ce grade de la maladie se détecte par l'apparition de néovaisseaux ; or, nos modèles de segmentation ne sont pas entraînés à détecter ceux-ci. Pour mieux quantifier le potentiel de l'approche par graphe, nous simplifions donc le problème posé, en regroupant les classes 0 et 1 d'une part (non malade et bénigne) et les classes 2,3,4 d'autre

part (modérée, sévère et proliférative.) Ce choix correspond aux recommandations dans le cadre de campagne de dépistage de la RD. On obtient alors les performances reportées dans le tableau 4.11. Dans le meilleur des cas, nous obtenons un couple Sensibilité/Spécificité de détection de 96.57%/81.29% sur Aptos et de 52.77%/98.06% sur EyePacs. On peut s'étonner d'une telle différence de sensibilité entre les deux bases : elle s'explique notamment par le fait que la base EyePACS contient notoirement énormément d'images de mauvaise qualité. Sur celles-ci, le réseau de segmentation tend à sous-segmenter les lésions en raison de la faible lisibilité de l'image. Inversement, Aptos est plus homogène, on obtient donc de meilleurs résultats de segmentation dessus. Ce dernier point confirme donc la nécessité d'utiliser des images de bonne qualité dès lors qu'on utilise la segmentation pour classer les images de fond d'œil.

Analyse des performances sous-optimales de $\mathcal{M}[\mathcal{B}^{(I)}]$

Dans le tableau 4.10, les graphes issus des modèles entraînés sur IDRiD et dans une moindre mesure sur MESSIDOR se démarquent particulièrement des autres pour leurs mauvaises performances de classification. Or, jusqu'ici, rien ne permettait d'affirmer que ce modèle produisait de plus mauvaises segmentations que les autres. Considérant que la difficulté à juger de la pertinence clinique des métriques de segmentation était la principale motivation derrière notre travail de classification ; nous pouvons déjà conclure à cet égard sur leur inadéquation relative à la classification par graphes.

Cependant, il existe une façon très simple d'améliorer ces résultats : connaissant la tendance de $\mathcal{M}[\mathcal{B}^{(I)}]$ à sursegmenter les petites structures (du fait du style d'annotation de sa base d'entraînement), nous avons réentraîné le modèle Att-GNN sur le graphe de lésions issu de ce modèle, mais en filtrant cette fois les lésions de taille inférieure à un certain seuil t . Les résultats sont affichés sur la figure 4.13. Il y apparaît clairement que c'est la surabondance de nœuds dans le graphe qui cause la dégradation des performances. Dès lors, rien qu'en filtrant toutes les structures de moins de 64 pixels, le score de classification sur Aptos est triplé. Malgré cette technique, le modèle n'atteint jamais les performances obtenues avec $\mathcal{M}_S$; ce levier d'action reste donc limité.

Classification des images après conversion adversariale

Un autre point qui mérite notre attention concerne l'efficacité de classification avec le modèle $\mathcal{M}[\mathcal{B}^{(R)}]$. Ses performances ne surpassent pas celles de $\mathcal{M}[\mathcal{S}]$, mais elles s'en approchent malgré tout (voir les dépassent sur Aptos). Or, lors de l'analyse des résultats de segmentation, $\mathcal{M}[\mathcal{S}]$ se révélait bien supérieur à $\mathcal{M}[\mathcal{B}^{(R)}]$ (tableau 4.4). Cela suggère que malgré sa segmen-

TABLEAU 4.10 Performances de classification (gradation en 5 classes) de la rétinopathie diabétique) à partir du graphe de lésions issu des différents modèles de segmentation. Deux architectures de classification sont comparées. On mesure le coefficient κ de Cohen quadratique sur deux ensembles d'évaluation distincts.

(a) Score obtenu par le PointNet++				(b) Score obtenu par le Attn-GNN			
Graphe issu de :	κ Aptos	κ EyePACS	Moy.	Graphe issu de :	κ Aptos	κ EyePACS	Moy.
$\mathcal{M}[\mathcal{B}^{(I)}]$	0.345	0.240	0.293	$\mathcal{M}[\mathcal{B}^{(I)}]$	0.206	0.206	0.206
$\mathcal{M}[\mathcal{B}^{(M)}]$	0.409	0.412	0.411	$\mathcal{M}[\mathcal{B}^{(M)}]$	0.755	0.551	0.653
$\mathcal{M}[\mathcal{B}^{(D)}]$	0.782	0.474	0.628	$\mathcal{M}[\mathcal{B}^{(D)}]$	0.801	0.505	0.653
$\mathcal{M}[\mathcal{B}^{(R)}]$	0.791	0.535	0.663	$\mathcal{M}[\mathcal{B}^{(R)}]$	0.822	0.606	0.714
$\mathcal{M}[\mathcal{B}^{(F)}]$	0.554	0.281	0.418	$\mathcal{M}[\mathcal{B}^{(F)}]$	0.602	0.314	0.458
$\mathcal{M}_{\mathcal{S}}$	0.815	0.549	0.682	$\mathcal{M}_{\mathcal{S}}$	0.789	0.682	0.736

TABLEAU 4.11 Performances de classification binaire dans un contexte de détection de la rétinopathie diabétique.

GNN	Graphe issu de :	Exactitude		Sensibilité		Spécificité	
		Aptos	EyePacs	Aptos	EyePacs	Aptos	EyePacs
PointNet	$\mathcal{M}[\mathcal{B}^{(R)}]$	0.898	0.869	0.872	0.368	0.916	0.988
	$\mathcal{M}_{\mathcal{S}}$	0.899	0.871	0.954	0.401	0.861	0.983
Att-GNN	$\mathcal{M}[\mathcal{B}^{(R)}]$	0.906	0.875	0.925	0.422	0.893	0.983
	$\mathcal{M}_{\mathcal{S}}$	0.875	0.894	0.966	0.528	0.813	0.981

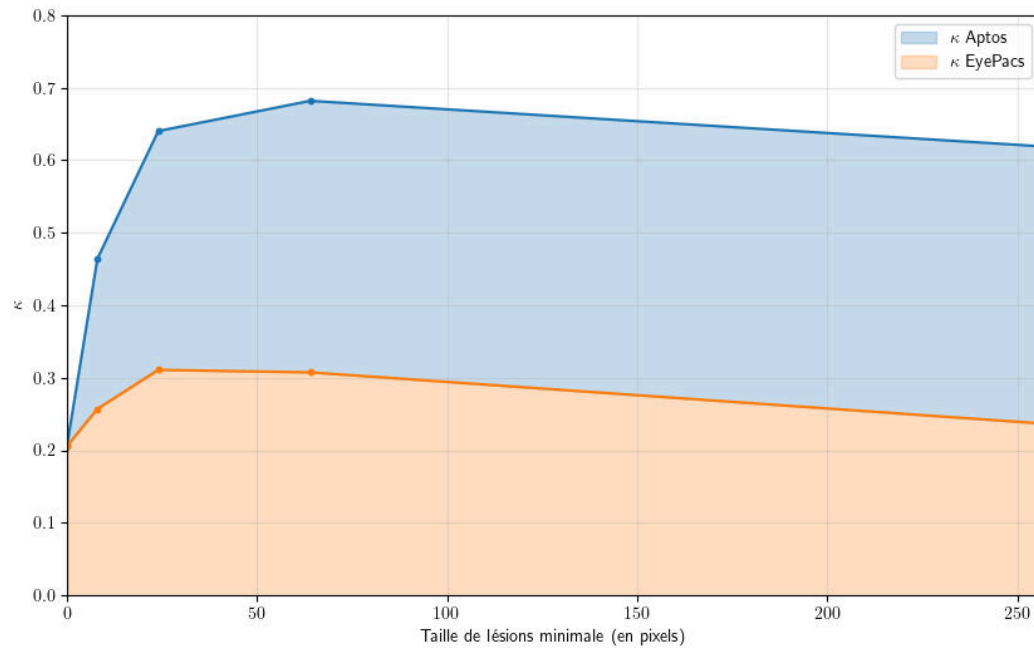


FIGURE 4.13 Effet du filtrage des plus petites structures segmentées par $\mathcal{M}[\mathcal{B}^{(I)}]$ sur la classification par graphe.

tation de moins bonne qualité, le style d’annotation du modèle entraîné sur la base $\mathcal{B}^{(R)}$ est propice à la classification. Or, rappelons qu’il s’agit du style produisant les segmentations les plus grossières (regroupant en de larges régions les structures voisines) : comme notre graphe est construit sur la base d’un nœud par structure connectée, une telle segmentation mène à une représentation par graphe bien plus allégée. Après avoir converti les images suivant notre schéma d’attaque adversariale, nous avons extrait les graphes et réentraîné les modèles. Les résultats sont synthétisés dans le tableau 4.12. Ils ne dévoilent pas de tendance très claire : les performances sur Aptos sont légèrement améliorées après la conversion vers RETLES, mais nettement détériorées sur EyePACS. En revanche, le modèle entraîné sur le graphe issu de $\mathcal{M}_S \Rightarrow I$ se révèle bien plus efficace que $\mathcal{M}_I$ (mais bien moins que $\mathcal{M}_S$) :

$$\begin{aligned} \kappa_{\mathcal{M}_R}^{Aptos} &> \kappa_{\mathcal{M}_S \Rightarrow R}^{Aptos} > \kappa_{\mathcal{M}_S}^{Aptos} > \kappa_{\mathcal{M}_S \Rightarrow I}^{Aptos} > \kappa_{\mathcal{M}_I}^{Aptos} \\ \kappa_{\mathcal{M}_I}^{EyePACS} &< \kappa_{\mathcal{M}_S \Rightarrow I}^{EyePACS} < \kappa_{\mathcal{M}_R}^{EyePACS} < \kappa_{\mathcal{M}_S \Rightarrow R}^{EyePACS} < \kappa_{\mathcal{M}_S}^{EyePACS} \end{aligned}$$

La conversion adversariale se comporte donc comme une forme de modèle intermédiaire entre le modèle généraliste et le spécialisé, observation qui est d’ailleurs en adéquation avec les résultats de segmentation. On a représenté sur la figure 4.14 trois exemples de graphes construits depuis les prédictions de $\mathcal{M}_S$, $\mathcal{M}_S \Rightarrow R$ et $\mathcal{M}_S \Rightarrow I$. Rappelons qu’il s’agit bien du même modèle, inchangé, prenant pour entrée la même image, mais avec différentes conversions adversariales (non perceptibles visuellement).

4.5 Conclusion

Nous pouvons résumer les principaux résultats de ce chapitre en dix points clés :

1. Nous proposons une caractérisation et une comparaison de bases de données publiquement disponibles, à la fois dans leur structure et dans leur style d’annotations.
2. L’architecture de segmentation sémantique développée obtient des performances de pair avec l’état de l’art. Nous n’observons pas de différences très marquées entre un modèle spécifiquement conçu pour le fond d’œil et des modèles de segmentation plus généralistes.
3. Même en maximisant un score sur une base de référence (type concours IDRID), le modèle ne maintient pas cette excellence sur d’autres bases. Une des raisons est que les styles d’annotations varient entre bases, les métriques mesurées ne sont donc pas comparables.
4. Une autre explication réside dans la variabilité des images et l’hétérogénéité des acquisitions en fonction des bases. Cependant, le prétraitement visant à modérer cet effet

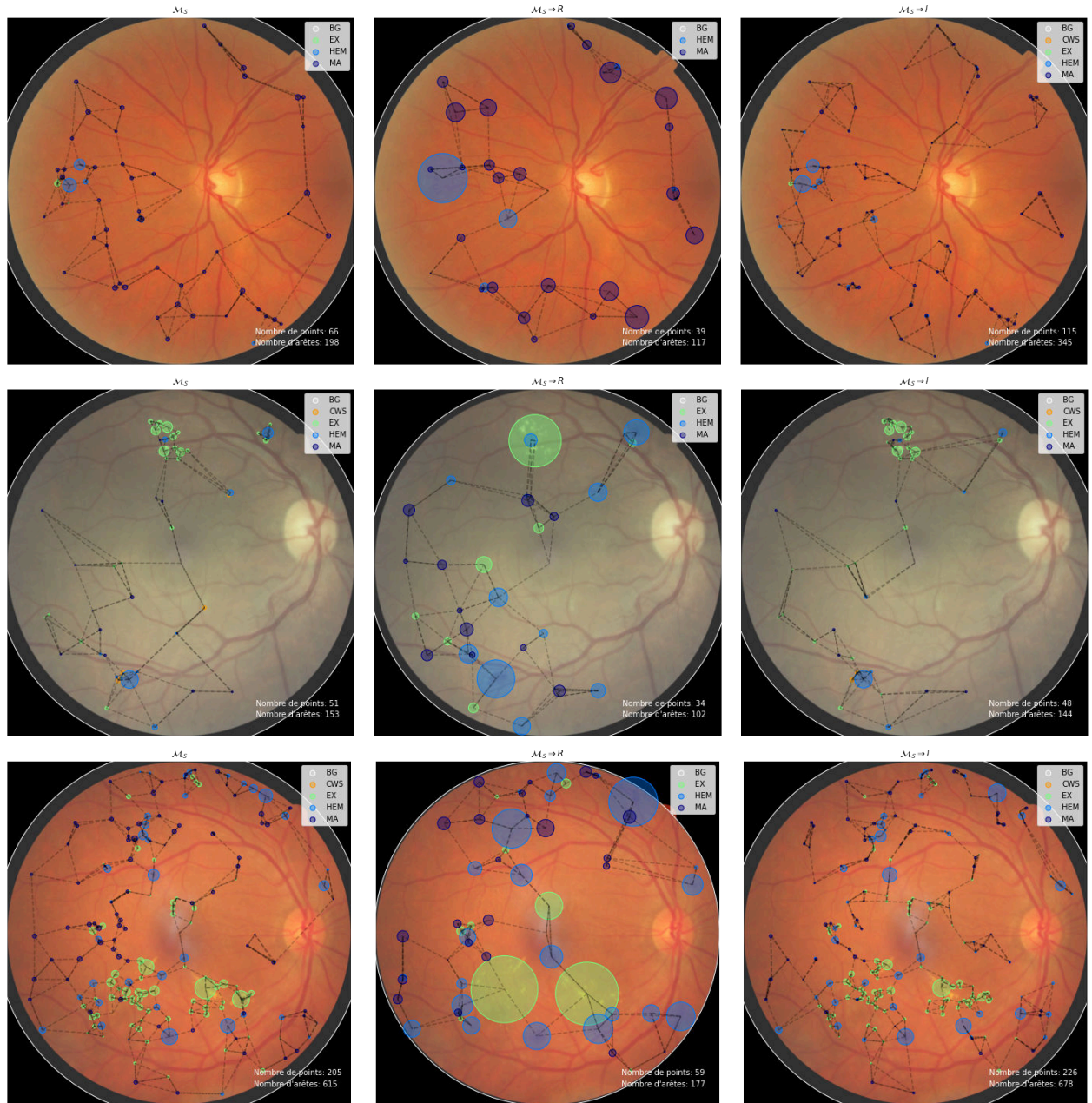


FIGURE 4.14 Exemples de graphes extraits à partir d'un même modèle M_S avec ou sans conversion adversariale des images vers IDRID ou RETLES. Sur ces images, chaque nœud est connectée à ses $k = 3$ plus proches voisins, mais elle est simplement indicative. Dans nos expériences de classification, $k = 24$.

TABLEAU 4.12 Performances obtenues après conversion adversariale vers le style d’annotation de RET-LES et IDRID. Pour cette dernière, nous donnons également les scores avec un graphe élagué des lésions de moins de 64 pixels.

(a) Gradation (classification multiclassées de R0 à R4) (b) Détection (classification R0-R1 vs R2-R3-R4)

Graphe issu de :	κ^{Aptos}	$\kappa^{EyePACS}$	Moyenne						
				Ex.		Sens.		Spéc.	
				Apt.	Eye.	Apt.	Eye.	Apt.	Eye.
$\mathcal{M}_S$	0.789	0.682	0.736	0.875	0.894	0.966	0.528	0.813	0.981
$\mathcal{M}_S \Rightarrow R$	0.795	0.613	0.704	0.888	0.877	0.956	0.483	0.841	0.970
$\mathcal{M}_S \Rightarrow I$	0.657	0.458	0.558	0.854	0.850	0.950	0.314	0.788	0.978
$\mathcal{M}_S \Rightarrow I^{(64)}$	0.731	0.476	0.604	0.866	0.856	0.961	0.337	0.801	0.979

n’améliore pas les modèles.

- Étonnamment, $\mathcal{M}_S$ parvient à maximiser simultanément les scores sur trois bases, aux annotations pourtant très distinctes. Il s’avère qu’il adopte un des styles d’annotations en fonction de l’image fournie : il est donc capable d’identifier son origine et d’adapter son style en conséquence.
- La question se pose donc de savoir comment : en disséquant $\mathcal{M}_S$ et en plaçant des sondes chargées de classifier la base d’origine de chaque image dans les caractéristiques calculées par le modèle, on observe qu’il est capable de « lire » l’origine de l’image.
- La sonde entraînée peut servir à générer des gradients dans la direction d’un style d’annotations au choix. Nous introduisons ainsi le concept de conversion adversariale d’une image : ce processus est stable, ne nécessite pas d’entraînement et ne modifie pas l’architecture visée. On observe que la conversion trompe non seulement la sonde, mais aussi le modèle de segmentation qui adoptera le style d’annotation de la base choisie.
- Un même modèle permet donc au choix de générer différents styles de segmentation. Il se pose alors la question de la pertinence diagnostique de chaque style. Nous proposons une méthodologie s’appuyant sur la construction d’un graphe rétinien.
- Nous obtenons des résultats de classification satisfaisants, mais loin d’atteindre les performances de l’état de l’art. En revanche, en tant que preuve de concept, les graphes présentent de nombreux intérêts : ils sont flexibles, leur entraînement est rapide et le modèle résultant est léger. De plus, il semble que dans cette approche, la qualité de la segmentation est corrélée avec la qualité de la classification. Cela facilite grandement son interprétation.

10. Si on exclut les performances de $\mathcal{M}_S$, il apparaît que RETINAL-LESIONS, base de données aux annotations les plus grossières, surpasse significativement les autres en termes de potentiel de classification (a fortiori sur des images de bonne qualité). Cela pourrait simplement indiquer qu'un graphe plus léger (en nombre de nœuds) est plus facile à classifier, mais cette hypothèse est contredite par les performances de $\mathcal{M}_S$ (qui génère bien plus de nœuds que $\mathcal{M}_R$). Cette observation permet d'enrichir la discussion sur la précision attendue lors des phases d'annotations manuelles : annoter grossièrement est significativement plus rapide et moins coûteux que de le faire finement. Sans offrir de conclusion définitive et catégorique sur le sujet, notre étude semble indiquer que la pertinence d'une annotation fine et détaillée n'est pas étayée par rapport à un équivalent plus grossier (mais aussi plus facile à obtenir). En revanche, nous montrons que les deux peuvent se combiner, y compris de manière déséquilibrée et que se faisant, on peut contrôler le style d'annotation du modèle pour s'adapter aux besoins.

Initialement motivé par la nécessité d'avoir un modèle de segmentation des structures de l'imagerie de fond d'œil, ce travail a progressivement gagné en ampleur. Il a permis d'étoffer la compréhension du fonctionnement de la segmentation sémantique, en couvrant à la fois le sujet de l'adaptation de domaine interbases et la pertinence de la classification couplée à la segmentation dans le diagnostic rétinien. Sans revenir sur les points évoqués dans les deux sections précédentes, l'étude permet une prise de recul sur la pertinence des métriques de segmentation, dans la mesure où celles-ci ne permettent ni d'apprécier la pertinence clinique de la segmentation réalisée (la précision sur les frontières des structures ne semble pas indispensable du point de vue du diagnostic) ni une comparaison interbase fiable. Ces deux aspects étant marginalement couverts dans la littérature, nous espérons inspirer des analyses plus approfondies des performances des futurs modèles qui seront proposés, analyses qui devront prendre en considération le questionnement sur la pertinence clinique.

L'intérêt d'une segmentation de qualité est reconnu et il en existe de nombreuses applications. Par exemple, une version préliminaire de nos travaux a déjà été utilisée par Peskine et al. [208] pour une évaluation interprétable de la qualité d'une image. Nous avons l'espoir que notre étude pourra, dans un futur proche, paver la voie vers une classification des maladies rétiniennes basée sur l'interprétation des structures sémantiques de l'image par une approche par graphe. Pour cela, de sérieuses améliorations des modèles sont encore nécessaires pour hisser leurs performances au niveau actuel des CNNs. Mais les concepts développés (conversion adversariale par sonde, représentation par graphe d'une image) ne sont pas spécifiques au fond d'œil et peuvent s'appliquer au-delà de notre seul domaine.

Si ce chapitre était consacré exclusivement aux fonds d'œil et dans une moindre mesure à la rétinopathie diabétique, le suivant élargira nos champs d'études à l'OCT et à d'autres

maladies.

CHAPITRE 5 INTERPRÉTABILITÉ D'UN MODÈLE DE CLASSIFICATION PAR ATTENTION CONCENTRÉE

Bien que partie intégrante du diagnostic des maladies rétinienne, l'utilisation de règles heuristiques ne s'appuyant que sur les lésions segmentées ne permet pas, en général, d'obtenir une classification automatisée robuste de la maladie. Ce constat est, dans une certaine mesure, confirmé par les résultats de notre précédent chapitre. Plusieurs éléments peuvent l'expliquer :

- Les algorithmes de segmentation ne sont pas assez fiables ; malgré leurs bonnes performances mesurées en comparaison avec une annotation manuelle produite par un expert.
- Les règles heuristiques définissant les étapes des maladies en fonction des lésions sont trop imprécises pour être strictement reproduites algorithmiquement.
- La définition de la maladie (ou de son grade dans le cas de la rétinopathie diabétique) varie selon le système de santé considéré. Par conséquent, les mêmes règles ne correspondent pas forcément au même stade pathologique.
- Lors du diagnostic, il existe une part d'interprétation subjective de l'image par le médecin qui use de son expérience et de son expertise.

Cette dernière hypothèse peut s'interpréter sous le prisme du formalisme Bayésien : l'ensemble des diagnostics possibles suit une certaine distribution de probabilités, la décision du clinicien consiste à choisir celui qui minimise son incertitude étant données ses connaissances a priori et de toute autre information dont il dispose (incluant les lésions qu'il identifie dans l'image). Par opposition, le suivi strict d'un ensemble de règles heuristiques s'appuyant uniquement sur les lésions serait équivalent à une classification par arbre décisionnel. En simplifiant grandement, la première appartient au monde de l'apprentissage probabiliste, là où la seconde est du ressort de la logique symbolique (c'est-à-dire régi par des règles formelles non-ambiguës).

Comme cela a été présenté dans la revue de littérature (Chapitre 3 et en particulier dans la section 3.1.1), l'essentiel des approches récentes sont statistiques et font usage d'une quantité massive de données. Elles permettent aux réseaux de neurones d'extraire d'eux-mêmes les marqueurs pathologiques via une représentation interne de l'image la plus souvent implicite. L'efficacité de ces approches est attestée dans de très nombreux cas, y compris pour des campagnes de dépistage de masse des principales maladies rétinienne systémiques. En particulier, les travaux de Bhaskaranand et al. [209], effectués sur plus de 100 000 visites de patients (soit 850 908 images) aboutissent à un système (désormais commercialisé) atteignant un couple sensibilité/spécificité de 91.3% / 91.1% pour la détection de la rétinopathie

diabétique. De telles performances (ainsi que le volume de données utilisé dans l'étude, non accessible publiquement) rendent extrêmement ardue la quête d'un système plus performant. Cependant, plusieurs travaux pointent les limites d'un tel modèle. Dans son éditorial pour la revue *Nature*, Styles [210] prend l'exemple du système de télédepistage écossais qui s'appuie depuis 2012 sur un système automatisé pour effectuer un premier filtre des images saines. Si la complémentarité et la collaboration entre le système informatique et l'expert humain sont désormais bien acquises, la performance de la machine ne peut se réduire à un couple sensibilité/spécificité. En effet, dans certains cas, il existera une ambivalence, voire une opposition, entre l'expert humain et la machine. Il convient par conséquent d'anticiper ces cas. Or, il est difficile de trancher en cas de désaccord si les mécanismes de prise de décision du modèle sont inconnus (ce qui est généralement le cas avec des réseaux de neurones pouvant contenir plusieurs millions de paramètres). L'un des axes de recherches qui permettent de résoudre ce problème vise à rendre interprétables les modèles employés, notamment a posteriori de leur entraînement. Un modèle cliniquement interprétable fournit à la fois un diagnostic précis et est également capable de souligner les structures locales de l'image qui l'ont guidé vers celui-ci.

Ayant ce contexte à l'esprit, l'objectif de notre travail de recherche est d'évaluer les performances d'un modèle récent de réseau de neurones reposant sur le principe de l'autoattention appelé *Vision Transformer*. Des performances élevées ont été obtenues dans plusieurs applications généralistes de reconnaissance d'images, mais parmi celles-ci, le domaine médical est encore peu exploré. Or, ces réseaux autoattentifs permettent structurellement l'émergence d'une forme d'interprétabilité pouvant identifier les segments locaux de l'image qui ont conduit à la décision du modèle, contribuant ainsi à rendre le modèle interprétable par le médecin. Un aperçu de cette propriété était déjà promu dans l'article original des ViT [6] et elle a par la suite été brièvement étudiée par Matsoukas et al. [211], qui s'interrogent sur l'opportunité de remplacer les CNNs par cette nouvelle architecture dans le domaine de l'imagerie médicale. Nous reprenons donc à notre compte l'hypothèse stipulant que la force (efficacité et acceptabilité) des réseaux autoattentifs découle structurellement de cette inhérente interprétabilité.

Dans ce chapitre, cette piste est explorée au travers de plusieurs scénarios mettant en jeu différents types de données ophtalmologiques. L'interprétabilité du modèle ainsi conçu est évaluée sous le prisme de l'acceptabilité clinique par un panel de rétinologues expérimentés. Pour tirer des conclusions générales, les expériences sont menées sur des données OCT et de fond d'œil, car ces modalités présentent chacune des difficultés spécifiques. Par ailleurs, pour approfondir la compréhension du fonctionnement interne du réseau autoattentif, nous proposons une étude du mécanisme d'extraction des patchs d'images et démontrons la pos-

sibilité d'améliorer la performance du modèle par ré-échantillonnage de ces patches. Cette idée, particulièrement simple à mettre en place (elle ne nécessite pas d'entraînement et peut s'appliquer à n'importe quel modèle existant) a inspiré une technique originale appelée « Attention Concentrée » (*Focused Attention*) visant à produire des cartes d'interprétabilité en s'appuyant sur un schéma d'échantillonnage conditionnel itératif permettant de filtrer progressivement les régions de l'image les plus pertinentes à sa classification. Pour valider à la fois notre hypothèse d'interprétabilité et d'efficacité des modèles type ViT par rapport à des CNNs, nous évaluons les performances des modèles sur la classification de bases d'images de fond d'œil et OCT, ainsi que sur la détection non supervisée de lésions. Nous avons par ailleurs mis en place une étude randomisée auprès d'un panel de quatre ophtalmologistes expérimentés afin de leur faire évaluer différentes méthodes d'attributions générant des cartes d'interprétabilité (dont l'Attention Concentrée).

Ainsi, pour résumer succinctement, les trois principales contributions présentées dans ce chapitre sont les suivantes :

1. Évaluer quantitativement les performances de l'architecture ViT et d'autres variantes et introduire un mécanisme de reparamétrisation afin d'améliorer la précision du modèle en phase d'inférence. Cette étude met en jeu plusieurs modèles, bases de données et modalités pour tirer des conclusions couvrant un large panel de configurations.
2. Proposer une nouvelle approche de la génération de carte d'interprétabilité appelée Attention Concentrée, reposant sur un échantillonnage conditionnel itératif des régions de l'image, donnant lieu à des cartes en hautes résolutions.
3. Valider quantitativement et qualitativement l'interprétabilité intrinsèque des ViT par rapport aux CNNs en proposant un comparatif entre plusieurs méthodes.

Ce chapitre est organisé de la sorte : la méthodologie employée est décrite dans la prochaine section 5.1, l'accent étant porté sur la notion de pas adaptatif et d'Attention Concentrée. Le protocole expérimental et les résultats sont détaillés dans la section qui suit 5.2. Nous en discutons la portée ainsi que les limites de la méthode en section 5.3.

5.1 Méthodologie

Ce travail repose sur l'architecture originale du ViT, dont le principe est décrit dans la section 3.1.2. Nos contributions -qui se greffent sur ce modèle- sont présentées dans les deux sous-sections qui suivent. Cette partie se conclut sur la description des stratégies d'entraînement, les jeux de données et les modules utilisés expérimentalement.

5.1.1 Affinage de l’attention par ajustement du pas

Nous proposons une méthode simple et qui peut s’étendre à n’importe quel type de modèle à base de jetons dérivé du ViT. Dans sa publication originale [6], le ViT décompose une image d’entrée en patches adjacents non superposés, projetés sur des vecteurs unidimensionnels et fournis au modèle sous la forme d’une séquence de jetons individuels. Cette opération est réalisée via une simple convolution dont la taille du noyau et le pas déterminent la taille de chaque patch et le chevauchement entre eux. Or, étant donné que la matrice d’attention est calculée à l’échelle des jetons, nous émettons l’hypothèse qu’augmenter leur nombre peut améliorer le modèle en fournissant des patches supplémentaires centrés sur davantage d’éléments structurels de l’image. Cette augmentation se fait tout simplement en diminuant le pas de la convolution projective. Ce faisant, on obtient une forme d’agrandissement de la résolution de la séquence, d’où le choix de parler d’échelle d’échantillonnage par la suite. Comme chaque élément de la séquence est accompagné de son encodage positionnel, à chaque échelle, de nouveaux encodages doivent être calculés pour les jetons suréchantillonnés. Si ceux-ci sont définis de manière formelle et analytique (sur la base de la position du patch dans l’image), leur calcul pour les nouveaux patches est simple. Pour nos expériences, nous avons fait le choix d’un encodage appris au cours de l’entraînement. Dans ce cas, l’obtention d’un nouveau vecteur d’encodage lors du suréchantillonnage de la séquence se fait simplement par interpolation des encodages existants entre ceux des jetons adjacents. Cette interpolation ne nécessite pas de réapprentissage, et de manière générale, l’ensemble de ce processus est conçu pour ne s’appliquer qu’à l’inférence sur un modèle préexistant, sans modification intrinsèque de ses poids. Néanmoins, la relation entre le nombre de jetons et la taille de la matrice d’attention est quadratique. C’est également le cas pour la relation entre le pas de la convolution projective et le nombre de jetons. En d’autres termes, une division par k du pas de convolution entraîne une augmentation par k^4 de la taille de la matrice d’attention à calculer. Cet effet a pour corollaire le risque d’une rapide saturation de la mémoire graphique nécessaire pour effectuer les calculs d’attention. Dans ces conditions, l’applicabilité de l’approche reste limitée au matériel haut de gamme, même pour une seule image et simplement à l’inférence (c’est-à-dire lorsque la rétropropagation des gradients n’est pas nécessaire). Ce fait justifie la poursuite de la recherche vers des techniques d’échantillonnage plus efficaces.

5.1.2 L’Attention Concentrée : un ré-échantillonnage itératif conditionnel

L’inconvénient du ré-échantillonnage naïf du nombre de jetons est lié aux limites matérielles sur la mémoire graphique disponible. La quantité de mémoire nécessaire augmente exponentiellement avec le facteur de division du pas de la convolution projective. Cette contrainte la

rend en l'état incompatible avec toute forme de technique qui requiert la rétropropagation des gradients, ce qui est notamment le cas pour l'essentiel des approches d'interprétabilité pour la génération de cartes d'attributions. On peut cependant arguer que tous les patches ne sont pas de la même importance pour la classification d'une image et a fortiori pour le diagnostic. Intuitivement, pour ce dernier cas de figure, les plus importants seront ceux centrés sur des lésions ou des marqueurs anormaux de l'image. La méthode proposée exploite cette hypothèse et consiste à générer une carte d'attribution et une prédiction en n'utilisant qu'un sous-ensemble des jetons pour chaque échelle (au sens donné dans la section précédente). Choisir ce sous-ensemble revient à attribuer un score d'importance par jeton et à filtrer les moins importants. Un tel score n'est naturellement pas trivial, mais c'est précisément le rôle que remplissent les algorithmes de générations de cartes d'attributions tels que l'*Attention Rollout* ou LRP. De tels algorithmes permettent d'indiquer les jetons de la séquence qui ont le plus contribué à la prédiction du modèle. Par extension, ils peuvent donc servir à définir des régions d'intérêts dans l'image, permettant un ré-échantillonnage des jetons dans la zone considérée.

Pour ce faire, on définit un processus itératif au travers d'une échelle de pas de convolution décroissants $s_i, i \in \{1, \dots, M\}$. Par simplification, on ne considère que des pas sous forme de puissance de deux, $s_i = \frac{s_0}{2^i}$, s_0 étant le pas du modèle originel. Notre procédure s'initie généralement avec $s_0 = 32$ ou $s_0 = 16$ et fait décroître s_i jusqu'à $s_M = 1$. En pratique, n'importe quel entier entre s_0 et 1 convient. À s_0 , le réseau est utilisé de manière classique, c'est-à-dire tel qu'il a été entraîné. Lors de cette étape d'initialisation, une carte d'attribution est générée via une méthode arbitrairement choisie, produisant un score pour chaque élément de la séquence de jetons. Celle-ci est ensuite suréchantillonnée (par la méthode du pas décroissant décrit dans la section précédente). À cette nouvelle échelle, seul un sous-ensemble de ces nouveaux jetons est fourni au réseau. La sélection de ceux-ci se fait par un échantillonnage aléatoire conditionnel basé sur le score d'attribution obtenu à l'échelle précédente. Le sous-ensemble est maintenu de taille constante P , de telle manière à contenir le calcul de la matrice d'attention à une taille $P + 1 \times P + 1$ (P jetons sélectionnés et un jeton de classification). Expérimentalement, nous avons aussi observé qu'ajouter les patches adjacents (dans l'image) à ceux sélectionnés par échantillonnage conditionnel améliore les résultats. Il est vraisemblable que ceux-ci permettent d'amener un contexte local autour des patches choisis.

La quantité P de jetons conservés devant rester modérée du fait des contraintes matérielles, nous répétons l'opération d'échantillonnage conditionnel K_i fois par échelle s_i , ce qui se traduit par K_i inférences par échelle. Cela a pour effet d'augmenter le nombre effectif de jetons considérés de $P \times K_i$ par échelle et par image tout en maintenant les exigences en mémoire

constantes. Étant donné que le nombre de jetons se densifie au fur et à mesure des échelles, on accroît le nombre d'étapes de ré-échantillonnage par échelle de manière progressive suivant l'équation $K_i = \lfloor 1.5^i \rfloor$ (où $\lfloor x \rfloor$ est l'arrondi à la plus petite valeur entière la plus proche de x). Les K_i cartes d'attributions produites par échelle sont fusionnées en une seule carte, elle-même agrégée aux précédentes cartes générées aux échelles inférieures. Nous avons expérimenté avec deux fonctions d'agrégation : la moyenne et le maximum élément-par-élément ; les résultats les plus probants étant obtenus avec la première. Pour chaque échelle s_i , la carte d'attribution est deux fois plus petite que celle produite à l'échelle suivante : si on souhaite s'en servir comme carte conditionnelle d'échantillonnage, il est donc nécessaire de la redimensionner, ce que nous faisons par interpolation bilinéaire. Le processus complet se résume donc à $\sum_{i=1}^M K_i$ inférences et à l'agrégation d'autant de cartes d'attributions.

Nous nommons cet algorithme *l'Attention Concentrée* : en effet, il vise à condenser le calcul de l'attention sur les zones de la séquence d'entrée les plus pertinentes. La figure 5.1 illustre le fonctionnement simplifié du processus. Le pseudo-code détaillé correspondant est fourni (Algorithme 1). Quelques exemples du processus d'agrégation de la carte d'attribution finale sont donnés sur la figure 5.2

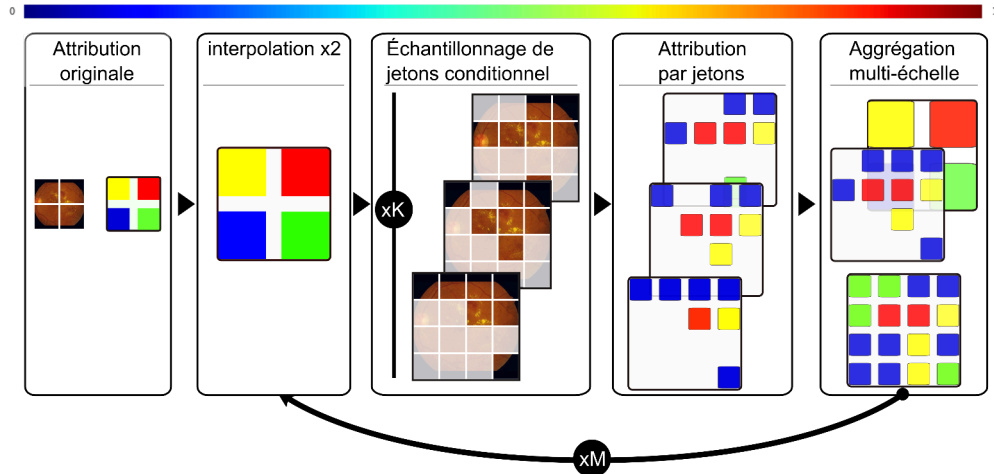


FIGURE 5.1 Mécanisme illustré de l'Attention Concentrée

5.1.3 Entraînement des modèles

Pour vérifier l'efficacité et la capacité de généralisation du ViT avec pas ajusté et Attention Concentrée, nous avons entraîné plusieurs modèles sur deux modalités d'imagerie distinctes. La grande majorité des éléments d'entraînements est conservée à l'identique entre les deux modalités ; néanmoins, certains hyperparamètres doivent être ajustés à chacune. Tous les

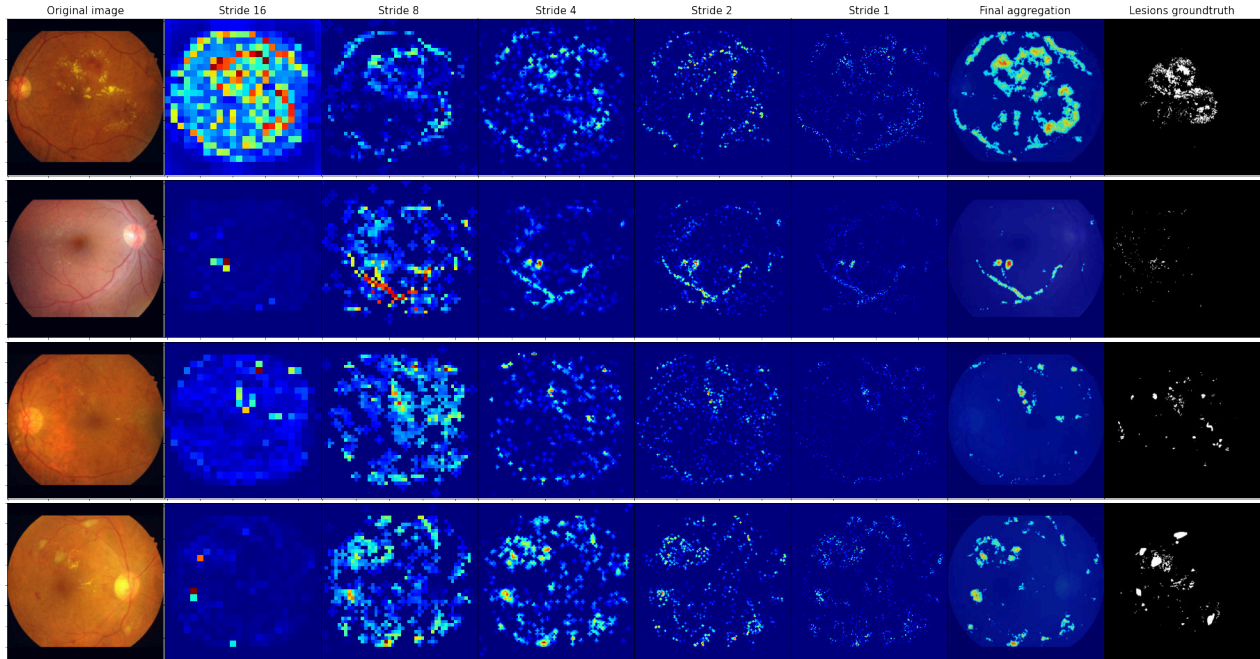


FIGURE 5.2 Exemples de la reconstruction de la carte d’attribution haute résolution par Attention Concentrée. Chaque colone représente un pas différent de l’opérateur de projection convolutif. L’avant dernière et la dernière colonnes représentent respectivement l’attribution finale et la vérité terrain des lésions contenues dans l’image.

modèles sont soit entièrement réimplémentés, soit dérivés des modèles fournis par la librairie de référence TIMM de R. Wightman [212].

OCT Les modèles sont entraînés pour minimiser la fonction d’entropie croisée avec lissage des labels (*label smoothing*) proposée par Szegedy et al. [213]. Une pondération par classe est ajoutée pour contrebalancer le déséquilibre de classe existant dans nos données. L’optimisation se fait par Descente de Gradients stochastiques avec une inertie (*Nesterov’s momentum*) de 0.9, un pas d’apprentissage de 0.0002 et une régularisation sur la norme L_2 des poids de 0.005 (*weight decay*). On utilise un planificateur de pas d’apprentissage, pour diviser celui-ci par 2 dès que le modèle atteint un plateau de performance (la *mean Intersection-over-Union*) sur l’ensemble de validation.

Fond d’œil Le pas d’apprentissage est fixé à 0.001. On augmente également le nombre d’itérations d’apprentissage. Hormis cela, les hyperparamètres sont les mêmes que pour l’OCT. De l’augmentation de données à la volée est employée grâce à la librairie Albumentations de Buslaev et al. [214] et consiste en des rotations et des mises à l’échelle aléatoires ainsi que des ajustements colorimétriques (luminosité, contraste, teinte...).

Algorithme 1 Pseudo-code pour l'Attention Concentrée

```

1:  $s \leftarrow 32$  ▷ Pas initial
2:  $P \leftarrow 256$  ▷ Nombre de jetons échantillonnés
3:  $attr \leftarrow obtenirCarteAttribution(model)$  ▷ Obtention de l'attribution originale (LRP, Rollout...)
4:  $i \leftarrow 0$ 
5: tant que  $s \geq 1$  faire
6:    $i \leftarrow i + 1$ 
7:    $s \leftarrow \frac{s}{2}$ 
8:    $model.stride \leftarrow s$ 
9:    $reechantillonnageEncodagePositionnel(model, s)$ 
10:   $attr \leftarrow interpolate(attr)$  ▷ Interpolation de la carte d'attribution
11:   $cdf = cumsum(attr)$  ▷ Obtention de fonction de répartition de la distribution d'échantillonnage des jetons
12:   $K = \lfloor 1.5^i \rfloor$ 
13:  pour  $k \leftarrow 0$  à  $K$  faire
14:     $u \leftarrow uniform(P)$  ▷ Échantillonnage d'un vecteur de taille P suivant une loi de distribution uniforme
15:     $indices \leftarrow searchsorted(cdf, u)$  ▷ Utilisation de la fonction de répartition conditionnelle pour choisir les jetons conditionnellement à l'attribution
16:     $voisins \leftarrow chercheVoisins(indices)$ 
17:     $indices.append(voisins)$ 
18:     $model.keptIndices \leftarrow indices$  ▷ Le modèle ne conserve que ces jetons
19:     $newAttr \leftarrow obtenirCarteAttribution(model)$ 
20:     $attr[indices] \leftarrow agregge(attr[indices], newAttr)$ 
21:  fin pour
22: fin tant que

```

5.1.4 Bases de données

OCT La majorité des expériences est menée sur la base de données UCSD OCT publiée par Kermany et al. [183] en 2018. Elle contient 108309 B-scans, rangés en quatre classes : Normal, présence de drusens, de néovascularisation choroïdale ou d'œdème diabétique maculaire. Parmi ce jeu, 1000 images ont été aléatoirement extraites pour être utilisées en tant qu'ensemble de validation. La base UCSD propose également un ensemble de test indépendant de 1000 images, parfaitement équilibré entre les quatre classes. Parallèlement, nous avons collecté 324 B-scans dont l'acquisition est faite à l'HMR de Montréal. Ces images ont été annotées par un expert de la rétine pour correspondre aux quatre mêmes classes présentes dans UCSD et ont servi comme ensemble de test complémentaire. Pour quantifier les performances des segmentations par carte d'attribution, nous utilisons la base de données AROI, publiée par Melinščak et al. [215] en 2021. Ce jeu contient 1136 B-scans, pris sur 24 patients

atteints de dégénérescence maculaire liée à l'âge au stade néovasculaire. Pour chaque image, des cartes de segmentation sont fournies, indiquant la présence des poches de fluides rétiens anormales (détachement de l'épithélium pigmentaire, liquide sous-rétinien et intrarétinien). Ces biomarqueurs sont en lien avec les quatre classes de la base de données UCSD. Par exemple, un détachement de l'épithélium peut être dû à une accumulation d'exsudats doux.

Fond d'œil Trois bases de données distinctes ont été utilisées pour évaluer notre méthodologie sur les images de fond d'œil. La base de données EyePACS, introduite par Cuadros et Bresnick en 2009 [203] et popularisée par une compétition organisée sur la plateforme Kaggle, contient un ensemble d'entraînement et d'évaluation de 35 126 et 53 576 images respectivement. Deux configurations ont été expérimentées :

- Apprentissage en faible régime de données, n'utilisant que l'ensemble d'entraînement.
- La combinaison des deux ensembles (pour un total de 88 702 images).

Pour cette deuxième configuration, comme il est bien nécessaire d'évaluer le modèle sur un ensemble séparé, une autre base de données est mise en jeu. Il s'agit de la base Aptos, également rendue publique lors d'une compétition Kaggle en 2019 [204]. Elle contient 3 662 images que nous n'avons utilisées que pour l'évaluation des modèles. Aptos et EyePACS sont de structures similaires, toutes deux contenant des images associées chacune aux cinq grades de rétinopathie diabétique (aucune, légère, modérée, sévère et proliférative). Cependant, les deux bases se distinguent par l'origine des patients, par la qualité des images et par la distribution des classes.

En supplément, nous avons également utilisé la base de données IDRiD [141], pour laquelle tous les détails sont donnés dans le chapitre 4. Pour toutes ces images, un même prétraitement est appliqué, consistant à centrer l'image sur sa région d'intérêt en retirant au maximum les bordures noires (voir la section 4.2.1 pour la méthodologie exacte). La résolution standardisée choisie est de 384×384 .

5.1.5 Architectures neuronales

En termes de modèles de référence, nous avons utilisé trois différents CNNs : le ResNet-152 [216], le Wide-ResNet-101 [189] et l'Optic-Net 71 [217] de Kamran et al. spécialement conçu pour la classification des images OCT. Ce dernier modèle n'ayant pas convergé sur la classification des images de fond d'œil, nous l'avons remplacé sur ces dernières par un EfficientNet B7. Le choix de ce modèle est motivé par ses performances au sommet de l'état de l'art sur la base de données ImageNet.

Dès lors ces références choisies, nous avons configuré nos expériences sur plusieurs modèles de type autoattentionnels (*Transformers*) en faisant varier leur architecture (ViT, DeiT, T2T-ViT) et leur nombre de paramètres. Dans les résultats fournis par la suite, ces modèles sont nommés suivant la nomenclature existante dans la librairie TIMM [212] (d'où sont également issus les poids préentraînés). Le suffixe « base » se réfère à un modèle utilisant un espace latent de taille 768, une profondeur de 12 couches et 12 têtes d'attention par couche. Le suffixe « large » indique respectivement des valeurs 1024, 24 et 16 pour ces mêmes dimensions. Par ailleurs, nous indiquons entre parenthèses la taille du pas initial de la convolution projective (plus le pas est grand, moins la séquence l'est). Enfin, pour le modèle T2T-ViT, l'espace latent est choisi de taille 512 et trois configurations sont expérimentées, correspondant à des profondeurs de 14, 19 et 24 couches.

5.1.6 Hiérarchisation des méthodes d'interprétabilité par attribution

Notre objectif premier n'est pas nécessairement de déterminer quel modèle est le plus performant, mais de savoir lequel est perçu comme pertinent dans un contexte clinique. Bien entendu, la pertinence est jugée à l'aune de sa performance de classification, mais également de sa capacité à fournir une prédiction interprétable, c'est-à-dire de justifier son diagnostic d'une manière ou d'une autre. Pour cela, nous avons fait le choix d'étudier les capacités d'interprétabilité visuelle (voir la section 3.3 pour un état de l'art sur le sujet). Plusieurs médecins ont été mis à contribution pour évaluer leur appréciation de différentes cartes d'attributions, générées par plusieurs modèles et en utilisant diverses méthodes d'attribution existantes et l'Attention Concentrée que nous proposons.

Pour mener à bien ce sondage, nous avons développé une plateforme accessible depuis un navigateur, permettant d'inspecter les images et leur carte d'interprétabilité respective. Une capture d'écran de l'outil est montrée sur la figure 5.3. L'outil permet de naviguer dans la base de données d'images, hébergée sur un serveur privé de l'École Polytechnique. Pour chaque expert, une connexion sécurisée est assurée par un service externe (*auth0*¹). L'ouverture d'une session sur la plateforme envoie une requête vers un serveur développé en Python avec le microframework Flask². Les images, leurs cartes d'attributions associées, les comptes des médecins et les annotations qu'ils réalisent sont stockés sur différentes tables SQL.

Pour une même image, les cartes d'attributions sont présentées selon un ordre aléatoire, sans préciser la nature de l'algorithme sous-jacent avant d'éviter de biaiser l'évaluateur humain. Il lui est possible de zoomer dans chaque image, d'agrandir l'interface et de contrôler l'opacité de la carte d'attribution superposée à l'image. Il associe à chaque carte un score compris

1. <https://auth0.com/>

2. <https://flask.palletsprojects.com/en/2.2.x/>

entre 1 et 5.

5.2 Protocole expérimental

Notre protocole est organisé comme suit : tous les modèles sont entraînés dans les mêmes conditions, afin de garantir une comparaison non biaisée et souligner ainsi leurs forces et faiblesses. Le même budget en termes d’itérations d’apprentissage est donc alloué à chaque modèle. Sur les deux modalités, les modèles sont testés sur plusieurs ensembles d’évaluation. En second lieu, nous évaluons l’impact du concept de pas adaptatif décrit dans la section 5.1.1. Enfin, la dernière étape est consacrée à l’évaluation des cartes d’attributions générées par divers algorithmes.

5.2.1 Performance de classification

OCT Afin de comparer les performances des différents modèles, nous avons calculé la précision, la sensibilité, la spécificité et l’exactitude sur les différents ensembles d’évaluation. Pour l’ensemble UCSD, les résultats sont donnés dans le tableau 5.1. Ils permettent d’illustrer la capacité des ViT à égaler les CNNs dès lors que le réseau est suffisamment complexe et utilisant des petits patches. Cela se paye en contrepartie par une augmentation drastique des ressources nécessaires pour faire fonctionner le modèle. Le modèle ViT le plus performant de notre échantillon est le ViT_{base}⁽¹⁶⁾, obtenant une exactitude de 97.80% (soit presque égale à celle du Wide ResNet101 de 98%). Cependant, ce modèle est significativement plus complexe que les autres en nombre d’opérations élémentaires requises pour une passe d’inférence. On note par ailleurs des performances en deçà de celles attendues pour les modèles de type T2T-ViT, mais également une complexité bien moindre.

En ce qui concerne la généralisation sur des données externes, les performances sur la base d’évaluation HMR sont bien moindres pour tous les modèles (voir le tableau 5.2). En moyenne, on observe une baisse de 7.51% tous modèles confondus. Dans le détail, elle est de 7.77% pour les CNNs, 7.97% pour les ViTs et de 6.63% pour les T2T-ViTs. Bien que cela puisse indiquer une meilleure capacité de généralisation pour ces derniers, étant donnée la taille limitée de la base de test et du nombre de modèles considérés, aucune conclusion ferme ne peut être tirée. Enfin, notons la relativement haute performance obtenue par le Optic-Net71, dépassant tous les modèles, y compris le ViT_{base}⁽¹⁶⁾ (deuxième de notre classement) de plus de 2%.

Fond d’œil. Comme décrit dans la section 5.1.4, nous avons entraîné les modèles suivant deux régimes de données (plus ou moins abondantes), utilisant 35 126 et 88 702 images res-

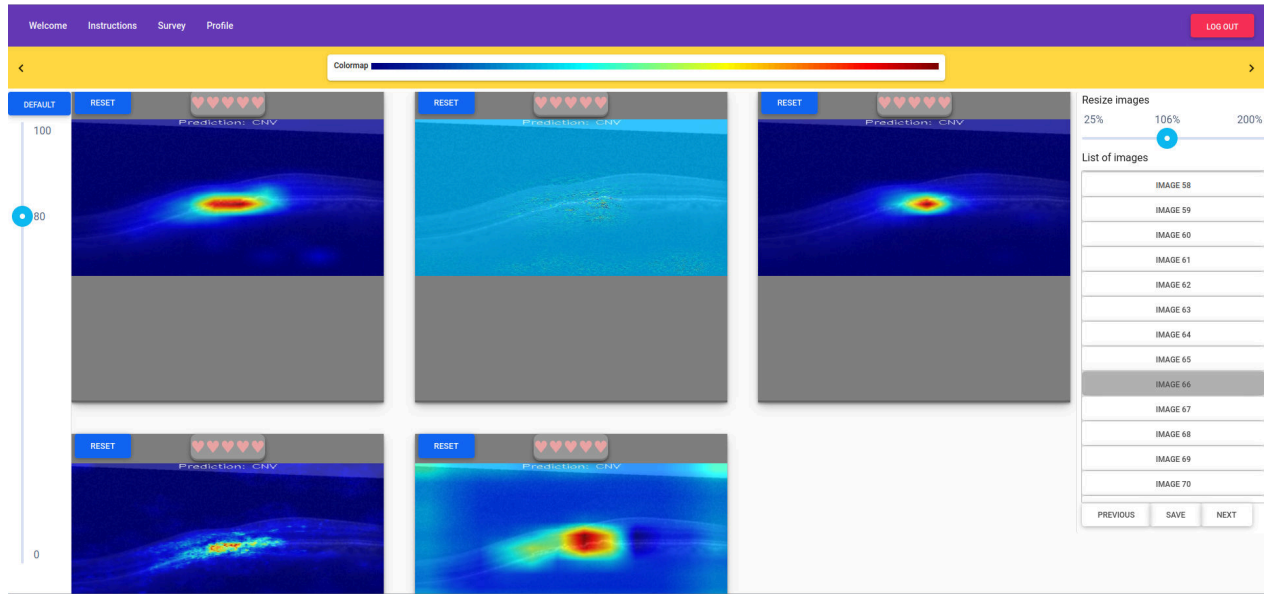


FIGURE 5.3 Capture d'écran de l'interface de la plateforme développée dans le cadre du sondage sur l'interprétabilité des réseaux neuronaux mené auprès d'experts médicaux.

TABLEAU 5.1 Performances comparées entre les modèles de référence de types CNNs et les modèles ViT pour la classification des OCTs de la base UCSD. Le nombre de paramètres (en millions) et le nombre approximatif d'opérations élémentaires d'addition/multiplications (*Multiply-Add Count* (MAC)) pour l'inférence sur une image sont donnés à titre indicatif sur la complexité des modèles.

Modèles	Exactitude	Spécificité	Précision	Sensibilité	Params. (M)	MACs (G)
Modèles de référence						
Wide Re- sNet101	98.00	99.33	98.15	98.00	477	124.85
Optic- Net71	96.80	98.93	97.06	96.80	48	12.45
ResNet152	96.40	98.80	96.61	96.40	223	58.15
Modèles auto-attentionnels						
ViT _{base} ⁽³²⁾	96.00	98.67	96.43	96.00	87	10.98
ViT _{base} ⁽¹⁶⁾	94.80	98.27	95.55	94.80	86	47.31
ViT _{large} ⁽³²⁾	90.20	96.73	92.40	90.20	303	162.0
ViT _{large} ⁽¹⁶⁾	97.80	99.27	97.93	97.80	303	162.0
DeiT _{base} ⁽¹⁶⁾	95.80	98.60	96.28	95.80	86	47.40
T2T-ViT ₁₄	94.40	98.13	94.91	94.40	21.5	6.1
T2T-ViT ₁₉	93.20	97.73	94.40	93.20	39.2	9.8
T2T-ViT ₂₄	93.40	97.80	94.67	93.40	64.1	15.0

TABLEAU 5.2 Performances comparées sur la classification des OCT de la base HMR.

Modèles	Exactitude	Spécificité	Précision	Sensibilité
Modèles de référence				
Wide ResNet101	89.51	96.62	91.39	88.49
Optic-Net71	92.59	97.62	93.22	91.96
ResNet152	85.80	95.40	89.19	84.42
Modèles auto-attentifs				
ViT _{base} ⁽³²⁾	87.65	95.95	90.07	86.21
ViT _{base} ⁽¹⁶⁾	88.27	96.15	89.44	87.31
ViT _{large} ⁽³²⁾	80.25	93.58	86.49	78.08
ViT _{large} ⁽¹⁶⁾	90.74	97.01	90.96	89.70
DeiT _{base} ⁽¹⁶⁾	82.72	94.42	87.22	81.24
T2T-ViT ₁₄	88.27	96.21	89.81	87.12
T2T-ViT ₁₉	85.80	95.47	89.18	84.19
T2T-ViT ₂₄	87.04	95.81	89.93	85.62

pectivement. Les résultats pour ces deux configurations sont donnés dans les tableaux 5.3a et 5.3b. La métrique utilisée est le coefficient Kappa (κ) de Cohen quadratique (voir la section 3.4) suivant une pratique courante pour ces bases de données. Sur cette modalité, on note les performances remarquables des variantes du T2T-ViT, dépassant significativement les autres modèles dans le régime à faible quantité de données. En revanche, les performances entre les différents modèles tendent à s'égaliser dans le second régime, confirmant que les ViT requièrent d'importantes quantités de données d'entraînement. Dans ce régime, on observe également que les trois variantes du T2T-ViT, le ViT_{base}⁽¹⁶⁾ et le ViT_{large}⁽¹⁶⁾ dépassent significativement les performances des modèles CNNs de référence. En moyenne, les modèles ont vu leur performance grimper de 1.53% entre les deux régimes. Dans le détail, cette augmentation est de 1.73%, 2.08% et 0.60% pour les CNNs, ViTs et T2T-ViTs respectivement.

5.2.2 Impact du pas adaptatif

Nous comparons ici les performances de différents modèles pour différents pas de la convolution projective, c'est-à-dire différentes tailles de séquences de jetons, suivant la méthodologie de la section 5.1.1. Plus ce pas est petit, plus longue est la séquence, conduisant à des exigences de mémoire GPU plus importantes. Nous avons pu diminuer le pas jusqu'à une taille de 3 pour une passe d'inférence sur une seule image en précision flottante (16-bits) sur une carte Nvidia RTX A6000. L'inférence est aussi significativement plus longue à cette échelle : il faut 30 millisecondes par image pour un pas de 16 contre 1.26 seconde pour un pas de 3. Les résultats pour l'OCT et le fond d'œil sont fournis sur la figure 5.4. Ainsi, en augmentant

TABLEAU 5.3 Performances comparées sur la classification des images de fond d’œil pour deux régimes de données.

(a) Régime basse-données (35 126 images).			(b) Régime haute-données (88 702 images).	
Modèles	EyePacs κ	Aptos κ	Modèles	Aptos κ
Modèles de référence			Modèles de référence	
Wide ResNet101	0.745	0.851	Wide ResNet101	0.879
ResNet152	0.743	0.864	ResNet152	0.876
EfficientNet B7	0.697	0.863	EfficientNet B7	0.875
Modèles auto-attentifs			Modèles auto-attentifs	
ViT _{base} ⁽³²⁾	0.658	0.852	ViT _{base} ⁽³²⁾	0.873
ViT _{base} ⁽¹⁶⁾	0.737	0.874	ViT _{base} ⁽¹⁶⁾	0.888
ViT _{large} ⁽³²⁾	0.605	0.836	ViT _{large} ⁽³²⁾	0.857
ViT _{large} ⁽¹⁶⁾	0.727	0.860	ViT _{large} ⁽¹⁶⁾	0.887
T2T-ViT ₁₄	0.746	0.886	T2T-ViT ₁₄	0.886
T2T-ViT ₁₉	0.743	0.878	T2T-ViT ₁₉	0.888
T2T-ViT ₂₄	0.752	0.878	T2T-ViT ₂₄	0.886

simplement le nombre de patches par image, on observe divers degrés de variation des performances, quelle que soit la variante du modèle considéré. Cependant, cette variation n’est pas monotonique : le pas optimal dépend du modèle et de la modalité d’imagerie et il n’apparaît pas de motif évident permettant de prédire le comportement des courbes. En revanche, nos résultats confirment que l’ajustement du pas est un hyperparamètre additionnel à prendre en compte dans l’affinage d’un modèle (sur un ensemble de données de validation). De plus, l’amélioration qui peut en découler est substantielle. En effet, sur le fond d’œil pour le modèle ViT_{base}⁽³²⁾, c’est 4% d’exactitude qui sont gagnés en diminuant le pas initial de 32 à 20. Nous souhaitons souligner que cette technique est utilisable en inférence sans requérir de réapprentissage du modèle ; elle est donc naturellement facile à déployer sur des modèles préexistants.

5.2.3 Évaluation clinique des qualités des cartes d’attribution

Rendre compte de la qualité et de l’interprétabilité d’une carte d’attribution générée par un modèle est une tâche relativement subjective. Dans un contexte clinique, cette évaluation nécessite une expertise médicale. En s’inspirant de la méthodologie de Singh et al. [218], nous avons demandé à quatre experts de la rétine de noter (sur une échelle de 1 à 5) la pertinence clinique de cartes d’attributions générées soit avec un CNN, soit avec un ViT_{base}⁽³²⁾. Les algorithmes que nous avons comparés sont les suivants :

- GradCAM (CNN)

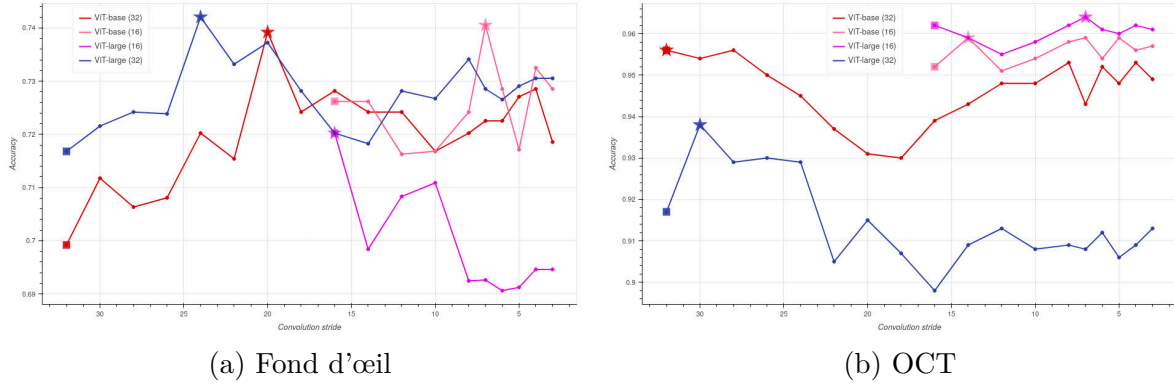


FIGURE 5.4 Effet du pas adaptatif sur les performances des modèles pour chaque modalité. Le point carré représente le pas initial du modèle (correspondant au pas lors de l'apprentissage). Le point en forme d'étoile indique le pas correspondant à une performance maximale.

- LRP (CNN)
- Attention Rollout (ViT)
- LRP (ViT)
- Attention Concentrée (ViT)

Le détail du fonctionnement de ces différentes approches est donné en section 3.3. L'Attention Concentrée n'est pas intrinsèquement une méthode d'attribution, mais a vocation à affiner une méthode préexistante. En l'occurrence, l'Attention Concentrée s'appuie sur l'algorithme LRP-ViT. Nous avons évalué la qualité des cartes d'attribution sur les deux modalités. Pour l'OCT, nous avons utilisé 150 images extraites de l'ensemble d'évaluation (50 images par maladie). Ces images ont été sélectionnées de telle sorte à sélectionner celles qui ont généré les cartes d'attributions les plus distinctes à travers les différents algorithmes (pour éviter aux cliniciens sondés d'avoir à évaluer des cartes trop ressemblantes). Pour le fond d'œil, nous avons choisi de mener notre étude sur l'ensemble d'évaluation de la base IDRiD, composé de 27 images.

Les résultats sont donnés sur la figure 5.5 pour les deux modalités. Pour les images OCT, l'Attention Rollout obtient le meilleur score moyen, dépassant légèrement l'Attention Concentrée. Dans le détail par maladie (Figure 5.6), on note que la préférence concerne notamment les attributions générées sur les DME (œdème maculaire diabétique). En ce qui concerne le fond d'œil, l'Attention Concentrée remporte très nettement l'assentiment des experts, confirmant ainsi la pertinence de la méthode. Pour vérifier que les scores sont significativement différents, nous avons utilisé un test de Friedman ($n = 177$ échantillons, $k = 5$ méthodes). L'hypothèse est vérifiée avec $p = 2.5 \times 10^{-7}$. En limitant la comparaison aux distributions correspondant aux trois méthodes utilisant un ViT (qui sont de facto les trois plus appréciées

par les experts), le test de Friedman confirme également une différence statistique entre les distributions $p = 0.0027$.

Pour compléter notre analyse des résultats de notre sondage, nous avons également mesuré l'accord interexpert. Étant donnés les quatre ophtalmologistes (A, B, C, D), nous avons calculé leur accord en utilisant le coefficient α de Krippendorff [128]. Nous obtenons un score $\alpha_{ABCD} = 0.41$. En regardant toutes les combinaisons possibles d'experts, nous obtenons un accord minimal entre A et D de $\alpha_{AD} = 0.23$ et maximal pour B et C, $\alpha_{BC} = 0.67$. Ces valeurs peuvent s'interpréter comme un accord allant de modéré à moyen entre les experts, ce qui met en exergue la subjectivité de la tâche. Notons cependant que nous mesurons ici l'accord sur la note attribuée et non sur le classement hiérarchique des méthodes ; l'accord étant probablement plus prononcé sur ce dernier.

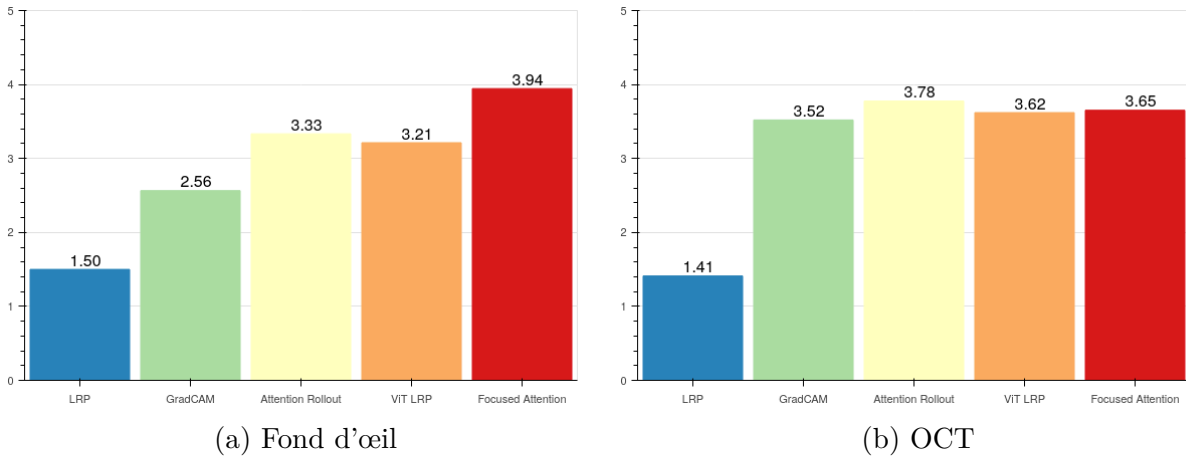


FIGURE 5.5 Scores attribués par les experts sondés concernant la qualité de différentes méthodes d'attribution. Les notes sont données de 1 à 5

5.2.4 Performance sur la segmentation des lésions

En complément du sondage clinique mené, nous avons également évalué la qualité des cartes d'attribution quantitativement. La disponibilité d'annotations de lésions sur les bases IDRiD (fond d'œil) et AROI (OCT) permet de mesurer les performances de segmentation obtenues en utilisant directement les cartes d'attributions. Nous avons également étoffé la gamme des algorithmes comparés. Pour les CNNs, il s'agit donc de GradCAM, GradCAM-guidé, les Gradients Intégrés et Deep Lift SHAP. Les implémentations de ces algorithmes sont fournies dans la librairie de référence Captum de Kokhlikyan et al. [219]. Plusieurs métriques sont utilisées pour évaluer la correspondance entre les annotations manuelles et les cartes d'attribution, dont l'AUROC et la Sensibilité-Précision Moyenne (*mean Average-Precision*)

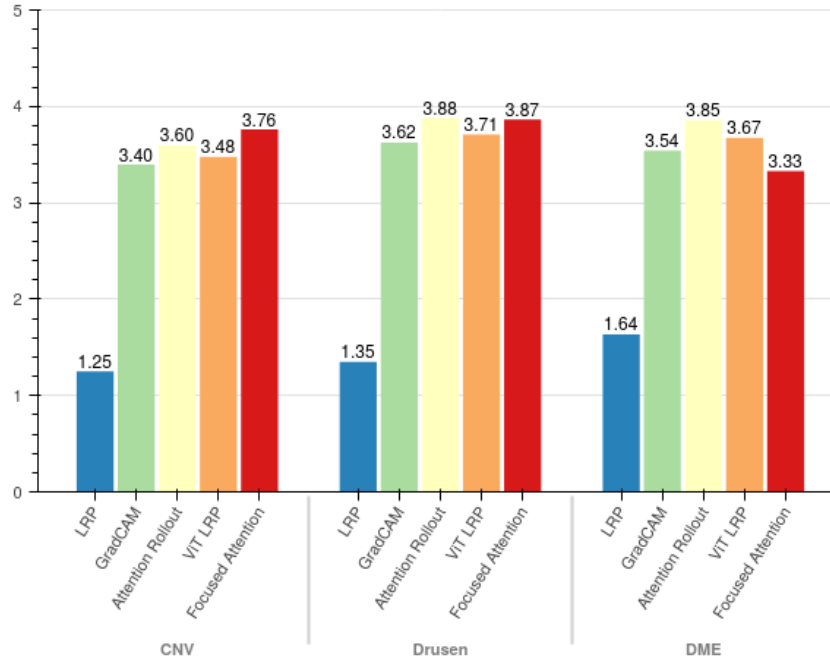


FIGURE 5.6 Scores d'interprétabilité sur l'OCT par maladie.

(mAP). Les cartes donnent une estimation continue par pixel, que nous avons binarisé avec un seuil de 0.5 pour le calcul de l'Intersection-sur-l'Union (IoU). Les résultats sont présentés dans les tableaux 5.4 pour les deux modalités. On y constate, en accord avec les résultats qualitatifs de notre sondage, que les méthodes issues du ViT présentent des performances très supérieures à ce qu'on obtient avec un CNN. De plus, sur le fond d'œil, l'Attention Concentrée permet un gain substantiel par rapport aux méthodes existantes. Elle génère également des cartes d'attributions de plus haute résolution par rapport à sa concurrence. Malgré l'observation faite en fin de section précédente concernant la subjectivité de la tâche donnée aux sondés, il est intéressant de constater que pour les deux modalités, le sondage et l'évaluation quantitative fournissent le même classement des méthodes.

Les figures 5.7 et 5.8 présentent des exemples qualitatifs de cartes d'attributions obtenues via les différents algorithmes, comparativement aux annotations manuelles de lésions fournies par les jeux de données.

5.3 Discussion

Notre étude souligne les nombreuses forces, mais aussi les faiblesses des modèles basés sur le principe du réseau autoattentif de type Transformer. Nous avons observé un net potentiel pour les ViT et variants à égaler voire dépasser les performances de classification d'un

TABEAU 5.4 Évaluation quantitative des performances de segmentation de lésions en utilisant les annotations manuelles disponibles pour les différentes modalités.

(a) Fond d'œil (Base IDRiD).				(b) OCT (Base AROI)			
Méthodes	mAP	AUROC	IoU	Méthodes	mAP	AUROC	IoU
CNN				CNN			
GradCAM	2.39	61.82	49.72	GradCAM	9.16	83.68	48.47
GradCAM-Guidé	1.61	46.25	49.47	GradCAM-Guidé	2.35	49.14	49.27
Gradients Intégrés	3.11	61.64	50.17	Gradients Intégrés	2.36	50.17	49.06
Deep Lift SHAP	1.79	56.73	49.24	Deep Lift SHAP	2.51	55.07	48.62
ViT				ViT			
Attention Rollout	11.21	77.81	52.83	Attention Rollout	25.06	94.47	57.48
LRP	11.58	81.27	52.80	LRP	19.19	93.55	55.83
Attention Concentrée	14.18	83.76	53.93	Attention Concentrée	19.15	92.89	52.25

modèle de type CNNs. Ce résultat est vérifié sur deux modalités très différentes entre elles, ce qui vient donner de la crédibilité à cette observation et permet de statuer sur la capacité d'adaptation de ce type d'architecture. En revanche, en termes d'inférence sur des images d'évaluation non issues de la même distribution que celles d'entraînement, notre expérience sur la base HMR témoigne que les ViT ne généralisent pas nécessairement mieux que les CNNs : dans notre cas, la dégradation des performances entre ces deux architectures est très similaire. Ce constat ne s'applique en revanche pas aux T2T-ViT qui parviennent à réduire cet écart de généralisation. En revanche, leurs performances sont globalement moindres, quelle que soit la base de données.

Un aspect qui semble légèrement distinguer le comportement des CNNs avec celui des ViT a trait à l'évolution des performances en fonction du nombre de données d'entraînement. En particulier, on a observé, sur le fond d'œil une plus grande capacité d'amélioration des performances des ViT par rapport aux CNNs en augmentant la taille de la base d'apprentissage (+ 2.8% pour les premiers contre +1.73% pour les seconds). Ce résultat encourage donc une mise à l'échelle plus grande et donc à la constitution de plus grandes bases de données. Quoiqu'il en soit, en l'état, nos expériences ne permettent pas de conclure sur une supériorité indiscutable d'une approche par rapport à l'autre, du moins en ce qui concerne les métriques d'évaluation communément utilisées.

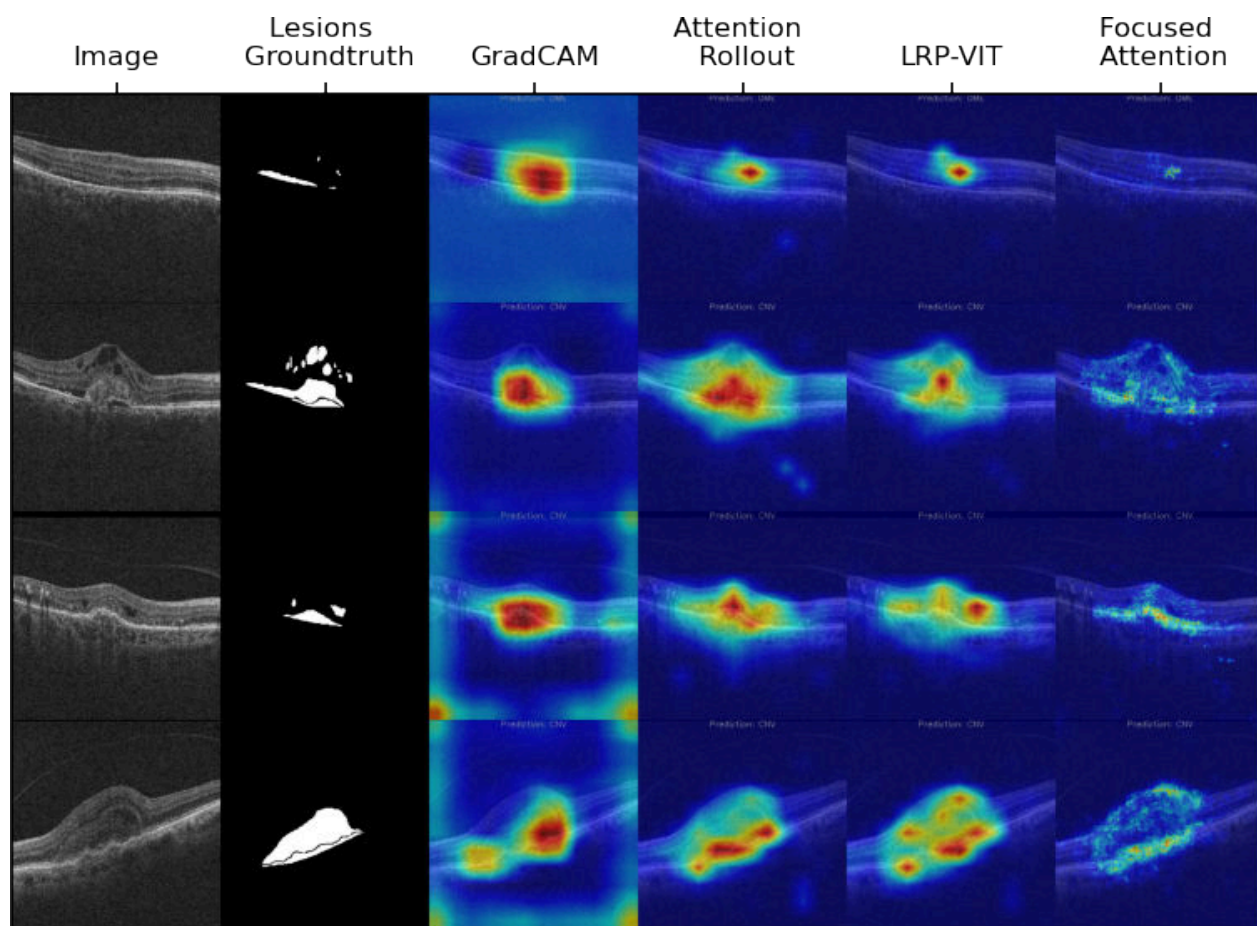


FIGURE 5.7 Analyse des cartes d'attributions issues de différents algorithmes sur des images OCT tirées de la base AROI.

En revanche, les ViTs se démarquent des CNNs sur l'interprétabilité de leur prédiction. Ce résultat est largement vérifié sur les deux modalités, quelle que la méthode utilisée pour générer les cartes d'attributions. Pour nous en assurer, nous avons à la fois interrogé des médecins sur leur appréciation des différentes cartes pour aboutir à des conclusions similaires.

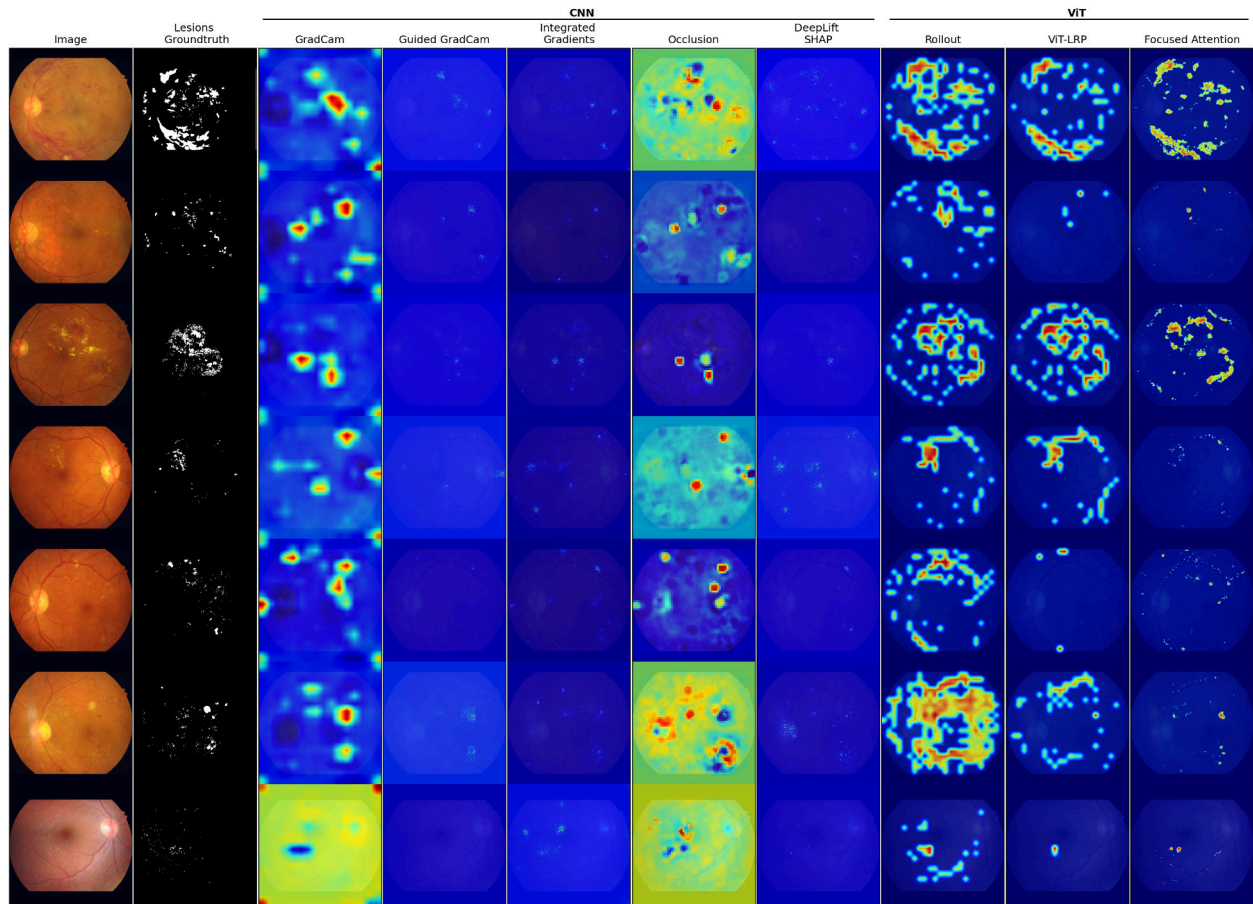


FIGURE 5.8 Analyse des cartes d'attributions issues de différents algorithmes sur des images de fond d'œil tirées de la base IDRID.

CHAPITRE 6 APPLICATIONS CLINIQUES DES ALGORITHMES DÉVELOPPÉS

Les deux précédents chapitres ont détaillé les méthodes propres à la généralisation et à l'interprétabilité que nous avons développées. Dans un cas comme dans l'autre, ce développement est globalement théorique ; dans leurs structure, application et validation, nos expériences répondent davantage au cahier des charges académique du champ de l'apprentissage machine plutôt qu'aux besoins cliniques. Ce chapitre propose de se détacher de cette logique : il y discute de plusieurs démarches exploratoires que nous avons entreprises pour créer une passerelle entre ces algorithmes et leur utilisation à des fins médicales. Les résultats obtenus étant parfois embryonnaires, cette présentation est avant tout un prétexte à une discussion plus approfondie sur nos travaux. Cela nous permet ainsi d'en illustrer les forces, mais aussi les limites ; d'où nous pouvons tirer un certain nombre de recommandations sur les perspectives de recherche.

Deux applications seront données en exemple à travers ce chapitre. La première porte sur l'identification du processus de conversion de la Dégénérescence Maculaire Liée à l'Âge (DMLA), du stade non exsudatif vers l'humide. Nos travaux sur la question sont une contribution à une étude menée par l'Hôpital Maisonneuve-Rosemont sur une cohorte de patients recrutée spécifiquement pour cela.

La seconde propose un prototype de déploiement d'algorithmes d'assistance au diagnostic à travers une plateforme de visualisation accessible en ligne. Nous y avons déployé nos différents algorithmes et une proposition de mise en forme de ceux-ci. Plus qu'une réelle application, il s'agit d'une démonstration d'un usage possible de ces algorithmes.

6.1 Recherche de marqueurs prédictifs de la conversion de la DMLA

La DMLA est une maladie largement répandue au sein de la population âgée de 50 ans et plus. Elle survient généralement sous deux formes : la plus fréquente étant la non exsudative (ou sèche), mais qui peut évoluer vers une forme exsudative (aussi appelée humide ou néovasculaire). La section 2.3.2 couvre brièvement les principaux marqueurs qui caractérisent un stade ou l'autre de la maladie. Nous renvoyons également vers les travaux de Spaide et al. [220], qui synthétisent la recherche d'un consensus international sur la terminologie derrière la classification de la DMLA et de ses symptômes. L'évolution de la maladie n'est pas symétrique entre les deux yeux ; en revanche le patient ayant développé une DMLA exsudative dans un œil a un risque significativement plus élevé de la développer dans le second [221, 222]. L'absence de

théorie sur le mécanisme de l'évolution de la maladie a motivé plusieurs études sur la question en particulier sur la prédictibilité algorithmique de la conversion de stade [223–225]. Dans ce contexte-là, l'équipe du Docteur Costantino de l'Hôpital Maisonneuve-Rosemont a mis en place un protocole de recrutement et de suivi de patients atteints de DMLA exsudative dans un seul œil (et donc à risque accru de conversion dans le second). Ces patients sont déjà suivis chaque mois dans le cadre du traitement de l'œil atteint de DMLA humide (injection intravitréenne d'anti-VEGF), mais dans le cadre de l'étude longitudinale, ils acceptent de se faire imager le second œil. La spécificité de l'étude repose sur la richesse des acquisitions : au cours de chaque visite, la rétine du patient est visualisée en utilisant :

- Une acquisition OCT structurelle prise avec un Spectralis de Heidelberg. Sa résolution est de $96 \times 1024 \times 436$.
- Deux acquisitions OCT structurelles prises avec un Zeiss PLEX, la première couvrant une zone de $3 \times 3\text{mm}^2$ (résolution : $300 \times 300 \times 1536$) et la seconde une zone de $6 \times 6\text{mm}^2$ (résolution : $500 \times 500 \times 1536$)
- Originalité de l'étude, le Zeiss PLEX nous fournit également des images OCT angiographiques, de même résolution que les structurelles.

Chaque acquisition est naturellement centrée sur la macula. Au début 2022, la cohorte s'établissait à 118 patients pour un total de 446 visites, soit moins de quatre visites par patient (3.77 en moyenne). Ce chiffre est très faible au regard de la date du début de l'étude (mars 2019). Il s'explique malheureusement par la pandémie de Covid-19 qui a occasionné une longue interruption dans le suivi de la cohorte.

Sur les 118 patients suivis, 13 (11%) ont développé une DMLA humide dans l'œil suivi.

6.1.1 Mises en application des algorithmes et expérimentations

L'objectif de notre participation à ce projet était de mener une étude préliminaire sur la prédictibilité de la conversion de la DMLA à l'aide des différents algorithmes à notre disposition, notamment ceux développés dans le cadre de cette thèse. Nous souhaitons mentionner ici les expériences conduites à titre d'illustration d'une application réelle, mais sans nécessairement rentrer dans le détail de l'implémentation ou des résultats obtenus. Cette précision a son importance, car ces travaux sont pour la plupart inachevés ou n'ont pas abouti à des résultats concluants. Cependant, même si les détails ne sont pas donnés ici par concision, l'essentiel des résultats et expérimentations est consultable sur le billet du blog conçu pour le suivi de ce projet : https://liv4d.github.io/Team_LIV4D.github.io/ophthalmology/oct/amd_progression_prediction/ ou en suivant le QR-code de la figure 6.1.



FIGURE 6.1 Lien vers le post de blog détaillant les résultats des expériences liées à la DMLA.

Classification des B-scans

Dans le chapitre 5, nous avons entraîné différents modèles à détecter différentes pathologies sur des B-scans d'une base de données publique (correspondant à des OCT structurales). Cette base contient les annotations de drusens et de néovascularisations choroïdales. L'une et l'autre sont respectivement des symptômes de la DMLA sèche et humide, les modèles sont donc a priori adaptés pour déceler une forme de progression de la maladie. Cependant, trois limites des modèles apparaissent immédiatement :

- Nos réseaux prédisent un diagnostic par B-scan et non par volume. On peut contourner ce problème en combinant la prédiction pour chaque B-scan du volume, ou encore en extrayant exclusivement le B-scan central (passant par la fovéa), ce qui correspond au scénario des données d'entraînement des modèles.
- Ils ne modélisent pas une évolution longitudinale, les prédictions ne sont donc pas corrélées dans le temps et par patient. Mais là encore, il existe plusieurs manières de combiner les prédictions par visite pour en tirer un score d'évolution (par exemple en faisant une moyenne temporelle glissante du score prédit).
- Enfin, nos modèles prédisent une classe discrète (éventuellement associée à une probabilité). L'extrapolation vers un score continu qui évoluerait au fil du temps n'est pas triviale. Une des possibilités est d'utiliser la probabilité associée à chaque classe comme indicateur.

Il faut aussi noter que les images acquises dans l'étude sont relativement différentes de celles qui composent la base publique UCSD [183] utilisée pour les entraînements de nos réseaux. La figure 6.3 donne une comparaison entre une image issue de celle-ci et des acquisitions d'une même rétine d'un patient de notre cohorte. À noter que sur UCSD, les acquisitions sont aussi faites avec un Spectralis, il est donc attendu que l'image 6.3b soit celle qui s'approche le plus de la distribution de cette base. Malheureusement, certains paramètres d'export de l'appareil n'étant pas à notre disposition (notamment le gamma de l'image et les valeurs

d'aplanissement de la rétine), les disparités restent importantes entre les images de la cohorte et celles d'UCSD. La question se pose donc sur la capacité de généralisation de nos modèles. Mais à défaut d'annotation sur ces images, nous évaluons la consistance des prédictions faites par le modèle pour chaque volume en fonction de l'appareil d'acquisition. Comme mentionné plus tôt, notre réseau fonctionne par B-scan plutôt que par volume, la question du mode d'agrégation se pose donc pour arriver à un score unique. Nous en avons expérimenté plusieurs : moyenne des diagnostics, identification par vote de la prédiction majoritaire au sein du volume, diagnostic restreint au B-scan fovéal ou sélection du pire diagnostic du volume. Pour chacune, on a mesuré l'accord entre les prédictions réalisées sur l'imagerie Zeiss $6 \times 6 \text{mm}^2$ et l'imagerie Spectralis (sur le même patient et à la même visite bien entendu). Force est de constater que cet accord est relativement faible (avec un coefficient κ de Cohen oscillant autour de 0.5 dépendamment de l'agrégation choisi), ce qui traduit à quel point le modèle est sensible au matériel d'acquisition des images. Cette expérience renforce donc d'autant plus (et sur un cas très concret) la nécessité de concevoir des modèles entraînés sur plusieurs appareils d'acquisitions. Il s'agit d'un cas très pratique où la conversion par sonde adversariale pourrait potentiellement améliorer le modèle. En s'appuyant sur nos résultats sur la généralisation de la segmentation de fond d'œil, une possibilité serait de placer une sonde chargée de détecter la modalité d'acquisition au cœur du réseau. Une expérience peut être conçue autour de cette idée : la sonde est entraînée à distinguer entre les images d'UCSD, nos images Spectralis et nos images Zeiss. Si cet entraînement converge, la conversion adversariale de l'image peut être tentée, avec comme critère de réussite un plus fort accord entre les prédictions réalisées sur le même patient avec les différentes modalités.

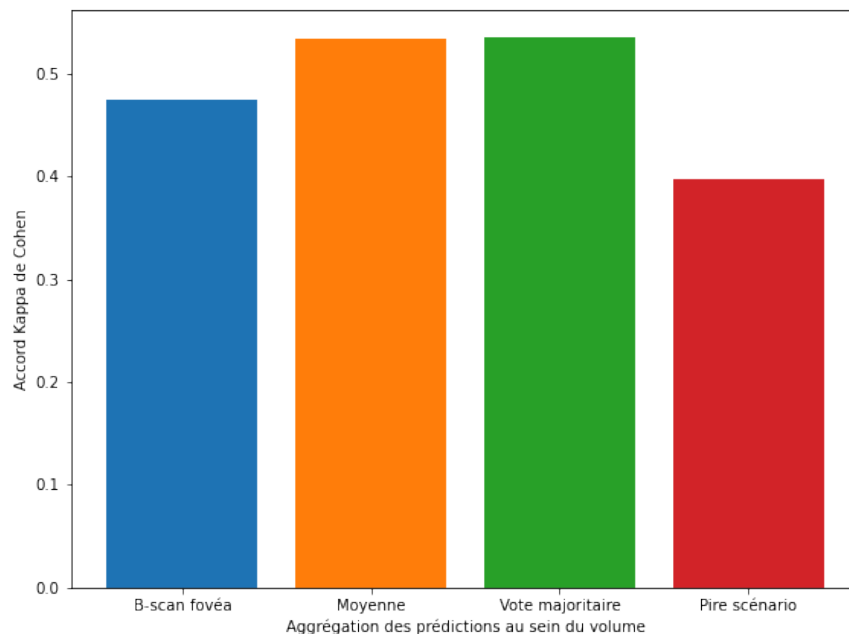
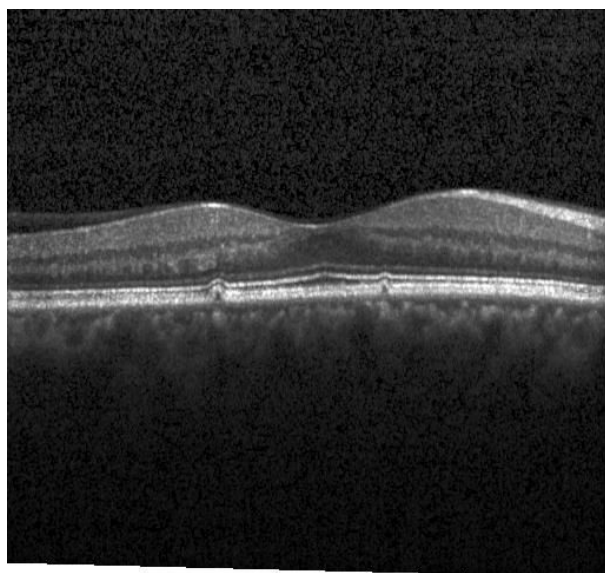


FIGURE 6.2 Accord au sens du κ quadratique de Cohen entre les prédictions réalisées par le même réseau sur les images Zeiss ($6 \times 6\text{mm}^2$) et Spectralis.

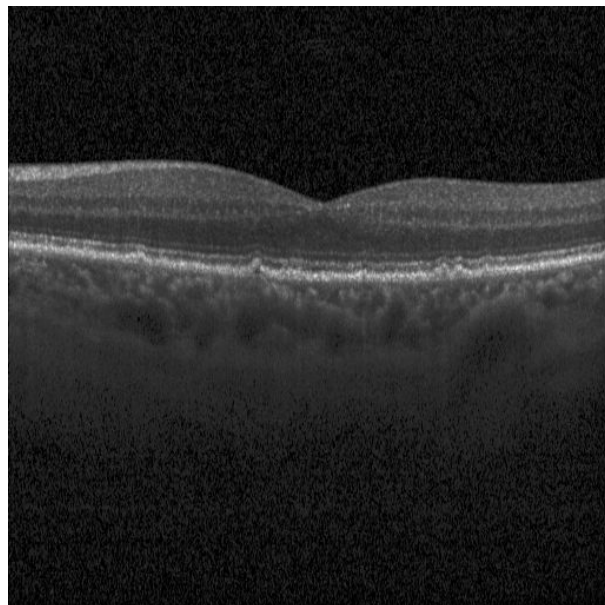
Quelles que soient les performances de généralisation, ajoutons que la prédiction en classes telles que « Drusen » ou « Néovascularisation » est une faiblesse conséquente du modèle, notamment car ces symptômes ne sont pas mutuellement exclusifs. Par ailleurs, un B-scan peut aussi bien posséder un seul que de nombreux drusens, ce qui change significativement le diagnostic. Une seule classe pour représenter cette diversité est donc malavisée, mais cette contrainte vient des données à notre disposition.

Utilisation des cartes d'interprétabilité

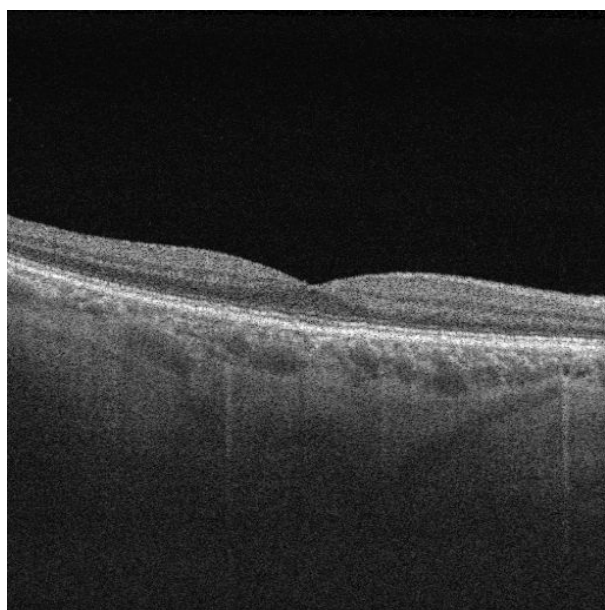
Une alternative à l'utilisation des prédictions de classification repose sur la segmentation des structures pathologiques et leur suivi au cours du temps. L'avantage est de pouvoir dériver plusieurs scores (taille des structures, leur nombre, etc.) qui corrigent les limites des prédictions par image de la section précédente. De plus, ces scores sont également plus faciles à interpréter cliniquement. Pour commencer, nous proposons d'utiliser les cartes d'attributions obtenues par les algorithmes d'interprétabilité, en particulier l'Attention Concentrée. Rappelons que celles-ci fournissent une indication approximative sur les structures locales qui ont guidé le diagnostic du modèle. La figure 6.4 fournit quelques exemples de drusens détectés par Attention Concentrée sur les patients de la cohorte. À chaque fois, nous avons extrait le B-scan central du volume Spectralis. Globalement, le modèle arrive à détecter correctement



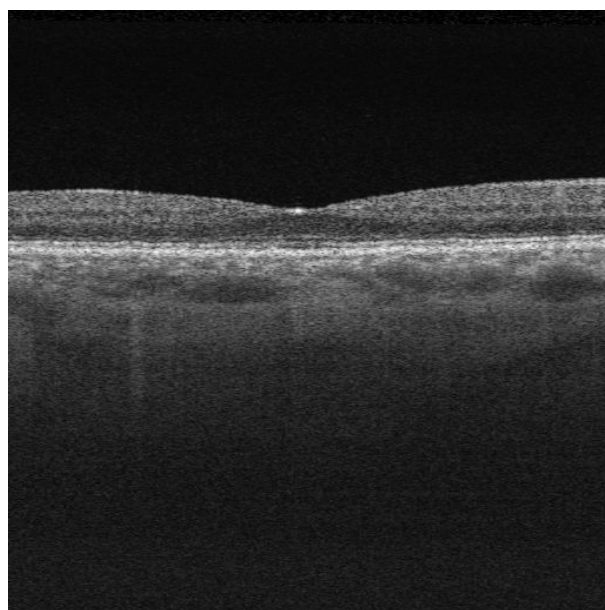
(a) Image de la base UCSD, classe « Drusen »



(b) Cohorte HMR : Spectralis



(c) Cohorte HMR : Zeiss $6 \times 6\text{mm}^2$



(d) Cohorte HMR : Zeiss $3 \times 3\text{mm}^2$

FIGURE 6.3 Acquisitions faites sur un même patient de la cohorte HMR comparées à un B-scan de la base publique UCSD sur laquelle sont entraînés nos modèles.

les drusens. Malheureusement, les indications fournies par la carte d'attribution restent peu précises : il ne s'agit pas d'une segmentation aux contours bien délimités qui permettrait de mesurer des statistiques de volume, largeur ou autre sur la lésion. Elles n'en restent pas moins utiles pour confirmer que le modèle n'hallucine pas lors de ses prédictions ; mais en l'état,

elles ne sont pas exploitables pour mesurer l'évolution des lésions. Pour s'en servir ainsi, une alternative possible est d'utiliser la carte d'attribution comme une forme d'initialisation ou de guidage pour un autre algorithme dédié à la segmentation et possiblement non supervisé, à l'instar d'un Watershed (Najman et Schmitt [226]).

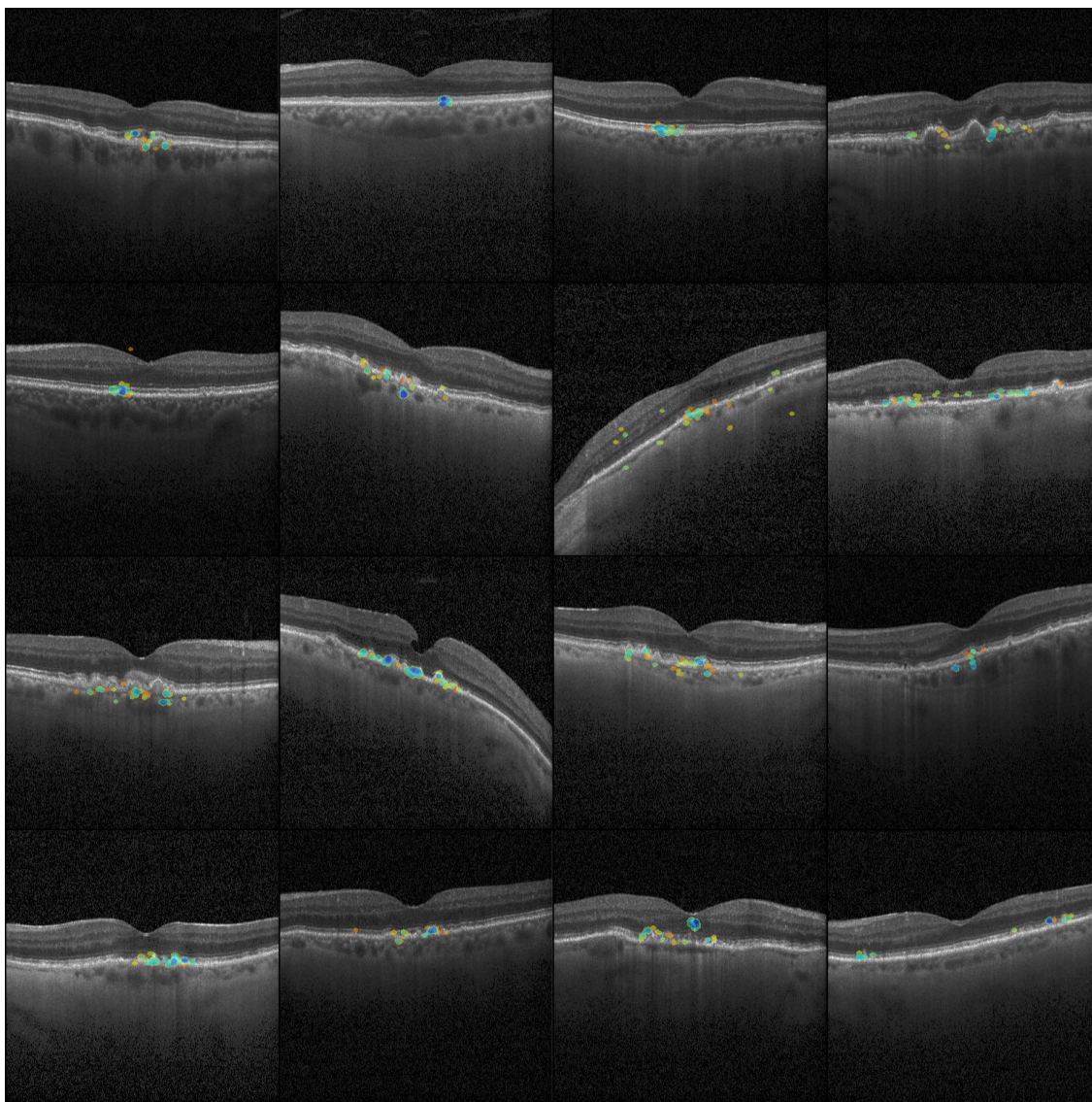


FIGURE 6.4 Attention Concentrée calculée sur les patients de la cohorte DMLA. Chaque B-scan correspond à un patient différent et est diagnostiqué comme présentant des drusens par le modèle ViT.

Classification de graphes de l'arbre vasculaire

En nous appuyant sur les succès relatifs de la classification de graphes rétiens avec un GNN, nous nous sommes interrogés sur la viabilité de cette piste pour la classification de nos données d'OCT angiographiques. En effet, l'angiographie fournit une visualisation précise des vaisseaux et néovaisseaux à travers les différentes couches de la rétine. En particulier, la projection de ceux-ci dans le plan de la surface de la rétine (appelée projection « en face ») fournit un arbre vasculaire similaire à celui qui est visible dans le fond d'œil (mais nettement plus détaillé). Pour cela, une première étape nécessite à identifier les différentes couches rétiennes, avant de pouvoir contrôler la profondeur de projection. Cette identification se fait donc via un premier modèle de segmentation. Là se ressent l'intérêt d'avoir les doubles modalités OCT structurelle et angiographique recalées entre elles : en effet, la segmentation des couches se fait dans la première et la projection de l'arbre vasculaire dans la seconde suivant la profondeur de la couche choisie. Nous avons donc entraîné un modèle de segmentation sur la base AROI publiée par Melinščak et al. [215]. Cette segmentation est illustrée sur la figure 6.5. Outre diverses lésions, nous obtenons surtout les interfaces délimitant les couches rétiennes.

À partir des cartes de segmentation de chacune des couches, nous pouvons obtenir les projections vasculaires à différentes profondeurs. Une illustration en est donnée sur la figure 6.6, représentant les acquisitions prises sur un même patient au cours de trois visites consécutives. À l'issue de la dernière, l'œil est diagnostiqué par un médecin comme présentant une forme de conversion (cependant le diagnostic n'est pas mené sur l'OCT angiographique, mais bien structurelle). Notre objectif est bien de retrouver sur ces images un marqueur d'évolution qui correspond à cette progression de la maladie. Mais il est notable au premier coup d'œil que les images ne présentent pas de variations brutales au cours du temps : si marqueur il y a, celui-ci sera très certainement subtil et probablement à chercher au niveau des capillaires. Sans rentrer dans les détails méthodologiques, nous proposons d'analyser l'évolution topologique de l'arbre vasculaire. Pour cela, la première étape est d'en réaliser une segmentation suivie de la squelettisation. Le squelette obtenu est propice à une représentation sous forme de graphe du réseau vasculaire : chaque vaisseau et ses embranchements y représentent un sous-graphe. Nous avons caractérisé chaque visite de patient par des métriques de graphes (nombres de croisements, d'arêtes, longueur de segments terminaux) dans l'espoir d'y déceler une tendance statistique distinguant nos deux populations (patient convertissant ou non), mais aucune caractéristique n'y est parvenue.

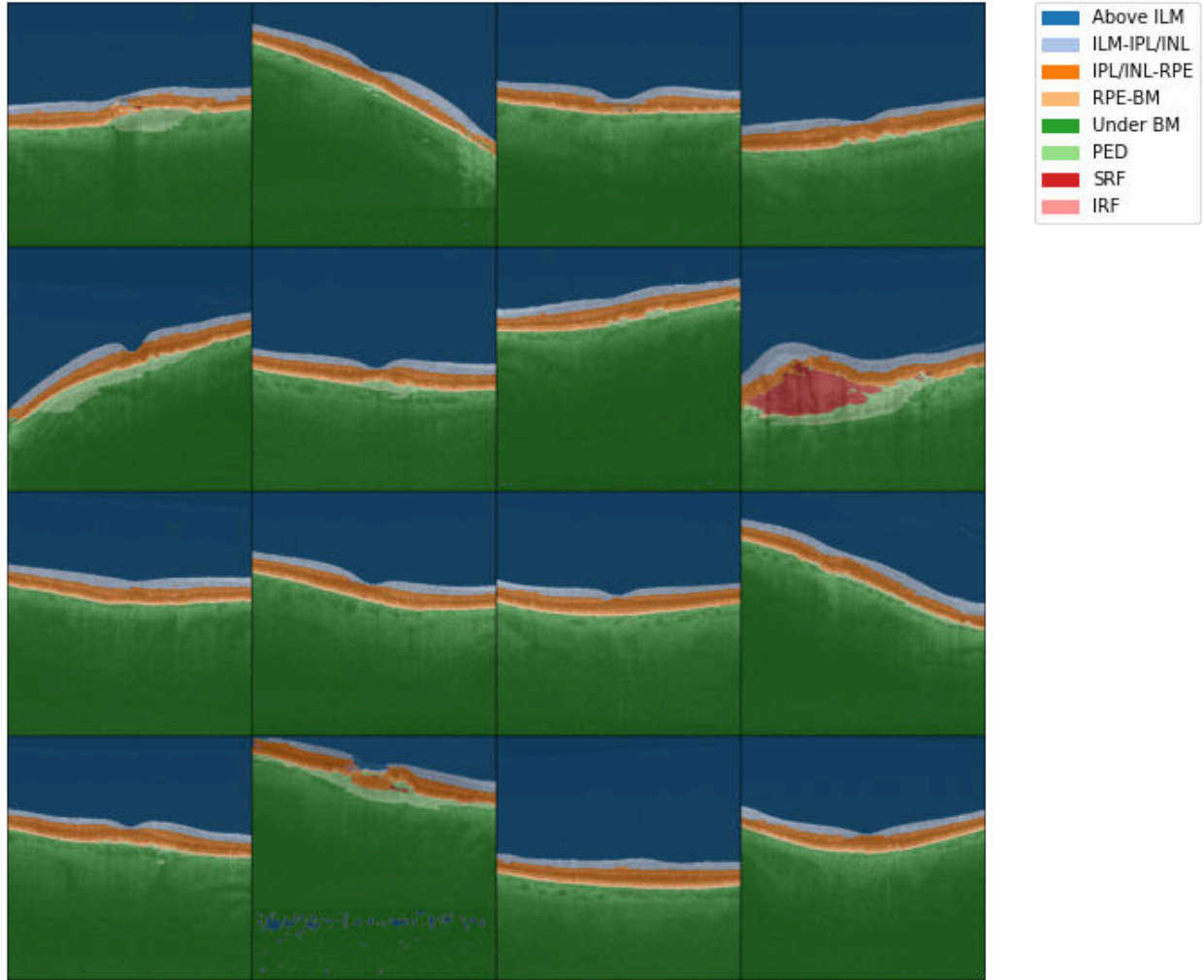


FIGURE 6.5 Segmentation automatique des couches rétinienne dans les images d'OCT structurales. Chaque image de la mosaïque est prise sur un patient différent. *ILM* : *Internal Limiting Membrane*, *IPL* : *Inner Plexiform Layer*, *INL* : *Inner Nuclear Layer*, *RPE* : *Retinal Pigment Epithelium*, *BM* : *Bruch Membrane*, *PED* : *Pigment Epithelial Detachment*, *SRF* : *Subretinal Fluid*, *IRF* : *Intraretinal Fluid*

6.2 Considérations sur un hypothétique déploiement clinique

La précédente section a couvert une application potentielle de nos algorithmes, orientée autour de la DMLA. Ces travaux ont pour objectif la compréhension de l'évolution de la maladie et ont un intérêt surtout théorique. Dans cette section, nous présentons une seconde application, qui porte sur le développement d'une plateforme d'aide au diagnostic clinique dans la pratique quotidienne. Cette plateforme peut aussi servir d'aide à la formation à la lecture d'images rétinienne. C'est un prototype (voire juste une simple démonstration de faisabilité)

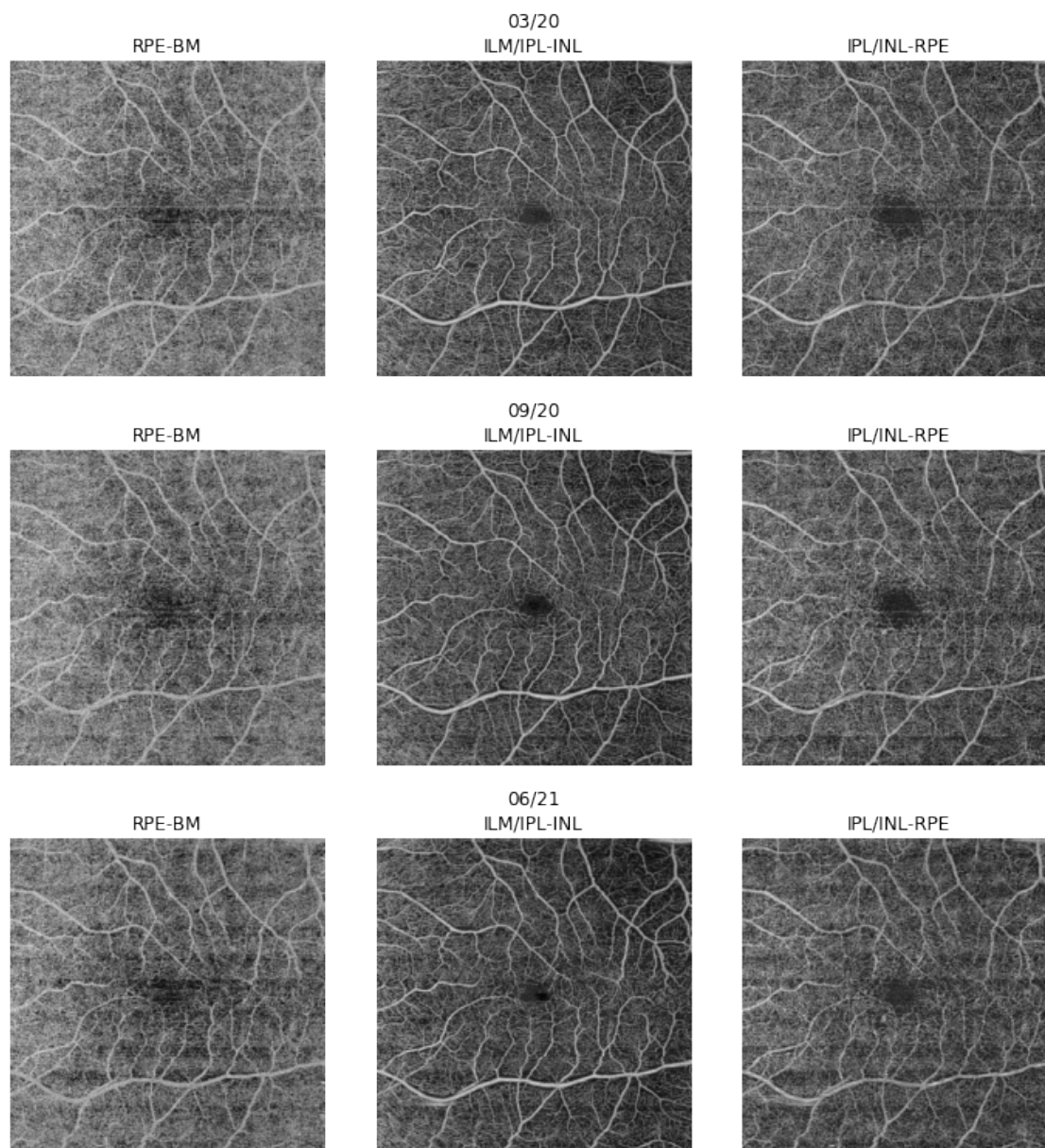
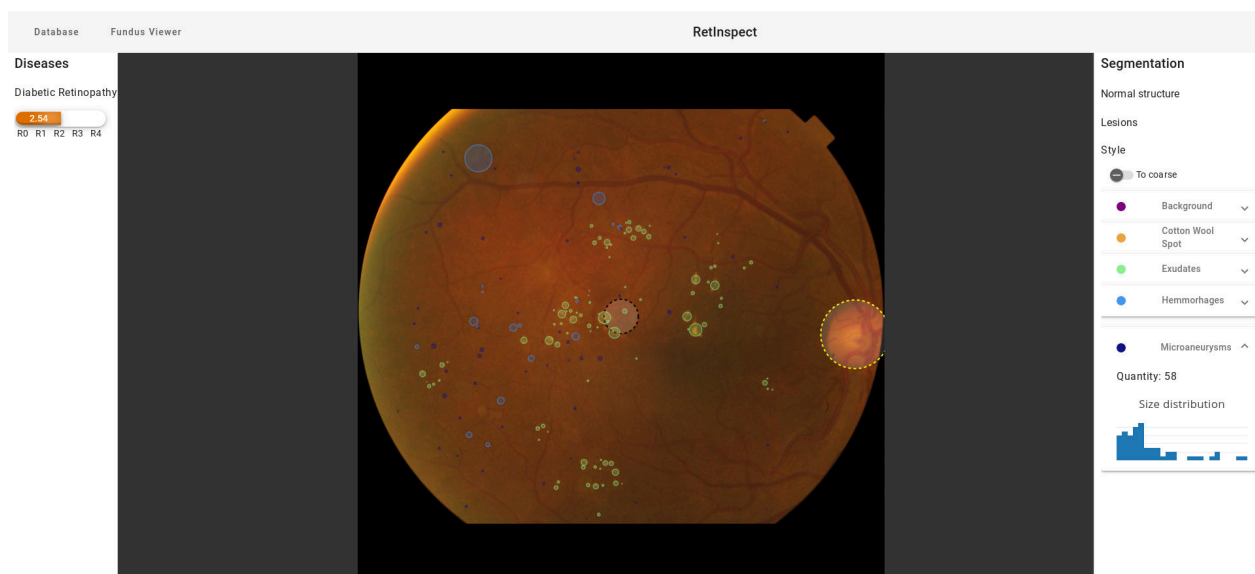


FIGURE 6.6 Évolution temporelle de l'arbre vasculaire d'un patient dont l'œil a subi une conversion vers une DMLA humide. Chaque projection vasculaire est réalisée à différentes profondeurs, obtenues grâce à la segmentation menée sur l'OCT structurale correspondante.

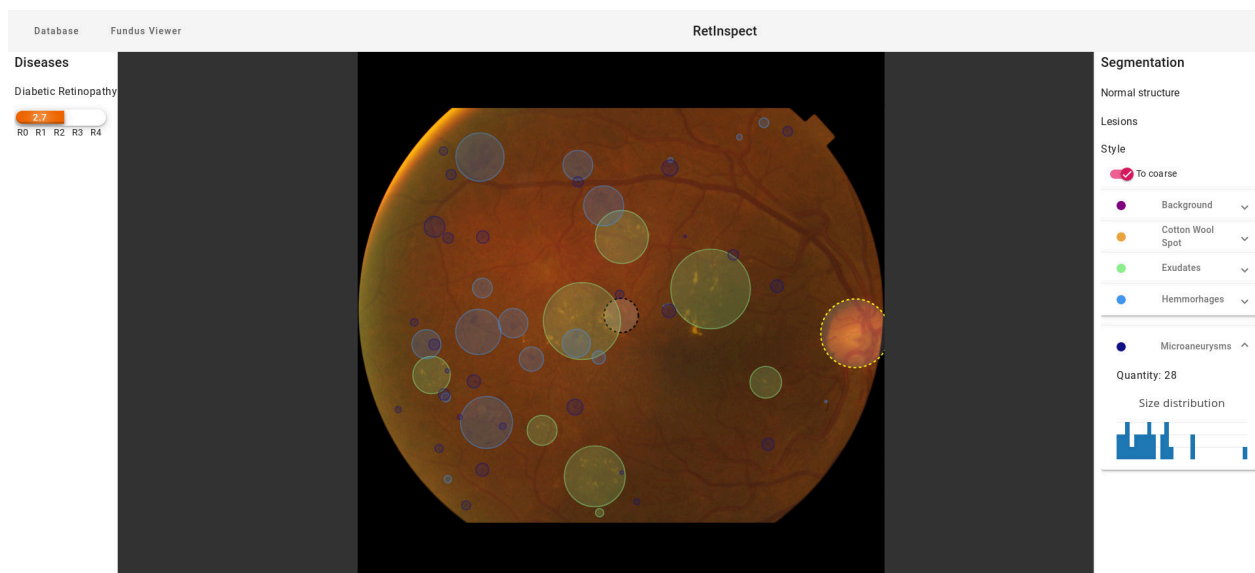
que nous introduisons ici, visant à démontrer un usage et l'intérêt des algorithmes développés. Sans rentrer dans les détails de son développement technique, la plateforme proposée fonctionne sur navigateur, elle ne nécessite donc pas d'installation. Elle permet à l'utilisateur de choisir une image de fond d'œil à traiter (ou de parcourir la base de données d'images existantes). L'image est envoyée sur un serveur, sur lequel les algorithmes de segmentation et de classification par graphe sont déployés. Le serveur renvoie alors la réponse des modèles,

comprenant le graphe rétinien et le score prédit de gradation de la rétinopathie diabétique. L'utilisateur peut alors visualiser ce que les modèles ont découvert dans l'image. Une double capture d'écran est présentée sur la figure 6.7 : les fonctionnalités de la plateforme sont relativement sommaires, mais elle permet d'apprécier l'intérêt pratique de la classification par graphes. En effet, cette approche se prête bien à la visualisation de différentes statistiques globales : quantité de chacune des lésions, distribution des tailles, etc. Un autre avantage de l'approche par graphe est sa rapidité d'exécution et le poids des modèles : comme nous l'avons vu, ces deux variables se comparent très favorablement aux CNNs conventionnels. Or, d'un point de vue technique, des modèles particulièrement légers permettent à terme un déploiement décentralisé directement dans le navigateur¹. Cet aspect est particulièrement intéressant au regard de la confidentialité des données, l'inférence dans le navigateur de l'utilisateur étant bien plus sécurisée que le transfert sur un serveur de calcul central. Bien que cet élément de discussion soit important, nous ne l'avons pas expérimenté sur notre plateforme. Pour être parfaitement complet, il faut aussi relever que même si le modèle de classification de graphes est léger, celui dédié à la segmentation ne l'est pas spécialement ; d'importantes optimisations restent donc à faire pour aboutir à une inférence complètement décentralisée. L'approche par graphe présente également un intérêt en terme d'interprétabilité des modèles. En effet, bien que nous n'ayons que survolé cet aspect dans nos expériences, il est possible de quantifier l'importance de chacun des nœuds dans la prédiction du modèle, en utilisant peu ou prou les mêmes algorithmes d'attributions locales que pour un CNN. Intrinsèquement, attribuer un score d'importance à un nœud n'a d'utilité que si ce dernier est facilement visualisable. En ce sens, nous arguons que la notion d'interprétabilité ne peut pas se décorrérer de l'outil mis à disposition de l'utilisateur pour l'appréhender. Notre prototype de plateforme répond justement à cette problématique : l'importance de chaque nœud s'obtient simplement via une fenêtre qui apparaît en le survolant de la souris. La figure 6.8 illustre ce concept : en plus du score d'importance à la classification, d'autres statistiques sont fournies, comme la taille de la lésion considérée ou sa distance à la fovéa (indiquée en pixels ou en fraction de diamètre de disque optique). Cependant, nous notons qu'une étude reste à mener auprès de la communauté médicale pour déterminer quels indicateurs sont les plus pertinents et facilitent le diagnostic de l'image. Enfin, cette plateforme permet aussi de prendre la mesure des limites de nos travaux sur la segmentation du fond d'œil. Tout d'abord, une segmentation automatique du disque optique et de la macula est clairement manquante : elles sont faites ici manuellement. De manière générale, il semble clair que l'ajout de biomarqueurs plus précis -que ce soient d'autres classes de lésions ou des structures non pathologiques (arbre

1. En s'appuyant sur des technologies encore en développement telles que le standard interplateforme d'export de réseaux de neurones ONNX et le langage WASM.



(a) Style de segmentation fin (type IDRiD)



(b) Style de segmentation large (type RETINAL LESIONS)

FIGURE 6.7 Plateforme de visualisation des prédictions faites par nos modèles. L'utilisateur peut choisir de faire une conversion adversariale du style d'annotations, ce qui influe sur le diagnostic du modèle.

vasculaire, nerf optique, fovéa)- devrait affiner la représentation par graphe de la rétine et potentiellement améliorer sa classification. Au-delà des structures locales manquantes, nous pouvons également ajouter que la classification à la seule rétinopathie diabétique est une limitation importante quant à l'intérêt des algorithmes. Bien qu'encore peu nombreuses, plusieurs recherches démontrent l'intérêt d'étendre la classification à d'autres maladies et des

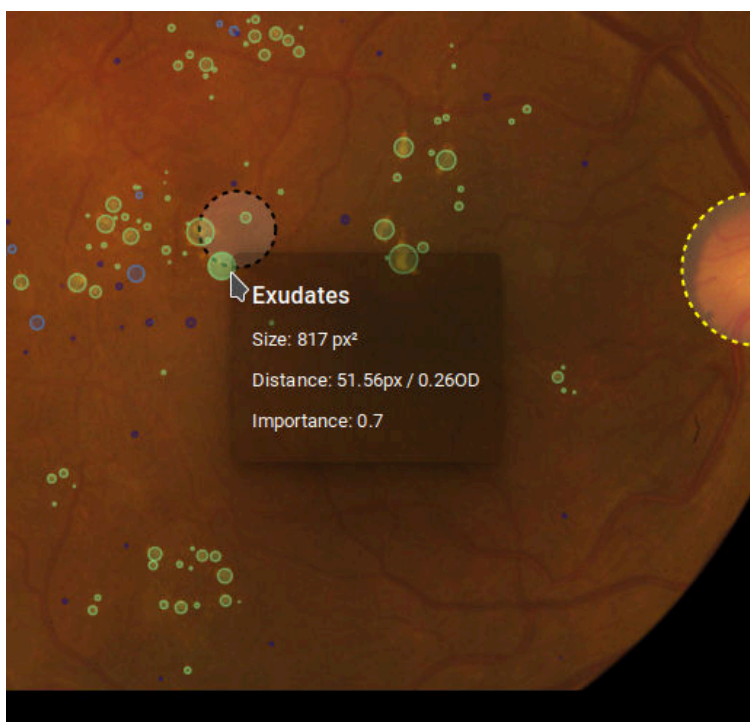


FIGURE 6.8 Illustration de l'importance d'un nœud dans la classification du graphe rétinien et autres statistiques associées consultables par l'utilisateur.

bases de données dédiées à cette tâche commencent à voir le jour (voir par exemple Panchal et al. [227]).

CHAPITRE 7 DISCUSSION GÉNÉRALE

Cette thèse a permis d’aborder des thèmes aussi variés que la segmentation sémantique de lésions, la représentation par graphe d’une image, sa classification et la conception de modèles de diagnostics autoattentifs mettant l’accent sur l’interprétabilité du réseau. Avant de détailler les contributions propres à ces différents segments, rappelons que notre ambition initiale était de concevoir un modèle interprétable de classification des pathologies rétinienne. Pour cela, nous avons expérimenté des méthodes diverses (CNN, GNN, ViT, combinaison de segmentation et classification...), toutes plus ou moins capables de répondre à l’objectif de la reconnaissance automatique de pathologies rétinienne. Dans ces expériences, la volonté n’était pas nécessairement d’obtenir un nouveau palier en matière de performances brutes de classification mais plutôt à proposer autant que possible une approche conceptuellement originale de la problématique identifiée. Les méthodes expérimentées (dont certaines peu explorées dans la littérature) présentent divers atouts et faiblesses ; en unifiant nos expériences au sein d’une même thèse, nous pouvons en dresser un inventaire comparé. Ce faisant, cela nous permet de conclure sur les jalons qui pavent le chemin vers un modèle unique, abouti et interprétable de classification des différentes pathologies rétinienne.

7.1 Limites des travaux menés

Pour chacune de nos contributions, nous proposons une nouvelle synthèse, mais écrite cette fois sous l’angle de leurs limites. Cet angle d’approche permet par contraste d’élaborer de nouvelles hypothèses à explorer pour tenter de dépasser les barrières actuelles.

7.1.1 Segmentation sémantique, adaptation de domaines et classification par graphes

Nous recensons ici les limites de l’étude du chapitre 4 portant sur la segmentation sémantique des lésions rétinienne.

Nous avons entraîné plusieurs modèles, suivant l’origine des données d’entraînement, avant de combiner celles-ci -malgré leur apparente incompatibilité en matière de style d’annotations- pour entraîner un seul modèle $\mathcal{M}_S$. On pourra arguer que l’entraînement du modèle de segmentation $\mathcal{M}_S$, central dans nos expériences, est relativement simpliste. Elle repose sur une combinaison de toutes les données confondues, sans tentative de ré-équilibre ni de calibration. Mais en pratique, cette manière de faire est étonnamment efficace et offre un ré-

sultat difficile à surpasser (on retrouve ce constat notamment dans les travaux de Hal Daumé III [228]). En pratique, nous avons tenté des approches plus sophistiquées afin d'améliorer notre modèle généraliste $\mathcal{M}_S$ (celles-ci sont détaillées en annexe C.3), mais nous n'avons pas obtenu de gains substantiels justifiant de les intégrer dans nos résultats.

Si $\mathcal{M}_S$ est on ne peut plus classique, la conversion adversariale via une sonde est la contribution la plus significative de ce chapitre. Elle soulève d'intéressantes observations sur la manière qu'a un modèle de générer son style de segmentation. L'utilisation d'une attaque adversariale non pas pour le tromper, mais pour l'adapter à un nouveau domaine en modifiant son entrée est à notre connaissance inédite. Elle permet de modifier la sortie d'un modèle sans le changer pour autant : elle ne requiert donc pas de changement de paradigme d'apprentissage. Cependant, plusieurs points limitent sa portée et seront à améliorer :

- La conversion adversariale reste coûteuse en ressources, car elle nécessite une rétropropagation itérative des gradients à travers le modèle de segmentation jusqu'à l'image. Bien entendu, cette opération peut se précalculer une fois par image et son résultat sauvegardé. Malheureusement, la modification de l'image est imperceptible, car d'une très faible magnitude (0.125% de variation maximale de pixels dans notre cas) : une sauvegarde en format conventionnel (PNG, JPEG...) 8-bits efface toute trace de celle-ci. Les réseaux de neurones manipulant des données en 32-bits flottants, il nous a fallu précacher les données sous un encodage permettant de conserver la trace de la variation adversariale. Cette contrainte purement technique alourdit très significativement la taille de stockage nécessaire pour une base de données convertie (d'un facteur 4 environ). Une solution serait de quantifier puis sauvegarder séparément les gradients calculés.
- L'utilité clinique de la conversion de style n'est pas clairement établie : certes, elle présente l'avantage d'être contrôlable et prédictible, mais son intérêt reste à quantifier au regard du potentiel de diagnostic. Pour cela, il serait bienvenu d'intégrer le retour de cliniciens en leur présentant les segmentations obtenues avec différents styles d'annotation comme les différentes hypothèses d'un même modèle ; un sondage sur la pertinence de l'approche (tel que mis en place dans le chapitre 5) pourrait se conduire. Dans le chapitre 6, nous avons proposé un prototype d'outil correspondant à cette proposition.
- La conversion adversariale a pour intérêt sa simplicité et sa prédictibilité, à contre-pied des approches d'adaptation de domaines reposant sur l'apprentissage variationnel [229] ou sur des modèles génératifs [230]. Ces approches n'ont pas été éprouvées sur le fond d'œil ou même sur une situation multibases (>2) telle que la nôtre. Une implémentation prenant en compte ces spécificités et une comparaison avec notre

approche de la conversion de domaines serait bénéfique pour souligner les forces et les faiblesses de chaque méthode.

Enfin, nous avons étudié le problème de la classification assistée par segmentation et avons montré sur plusieurs cas pratiques de CNNs la difficulté de trancher sur la pertinence de cette dernière pour aider le diagnostic. Pour clarifier cette question, nous avons développé une approche alternative, basée sur la représentation par graphe des lésions rétinienues. Ce faisant, la classification du graphe ne peut faire l'économie des informations issues de la segmentation, ce qui nous permet ainsi indirectement de mesurer la qualité de celle-ci. Cependant, de toute évidence, les performances de la gradation de la rétinopathie diabétique avec le graphe sont significativement en deçà de celles obtenues avec un CNN. Elles s'approchent davantage des valeurs de la littérature quand on simplifie le problème à celui de la détection binaire de la maladie. Plusieurs pistes pourraient potentiellement améliorer ce type de modèle : la prise en compte d'un nombre plus varié de lésions (incluant les néovaisseaux), voire l'intégration des structures normales dans le graphe rétinien. Ces éléments peuvent d'ores et déjà être étudiés, car certaines bases de données fournissent ces annotations. Elles restent cependant encore rares, ce qui induit des contraintes supplémentaires sur l'apprentissage en faible régime de données.

Un des avantages de la classification par graphe est son extrême légèreté et sa vitesse d'entraînement. Avec l'Attention-GNN (1,4 million de paramètres), une époque sur 35126 graphes prend environ 16 secondes d'entraînement ; à comparer aux 57 minutes en moyenne (sur autant d'images) que prennent les modèles CNNs de la section 4.4.1 (45.7 millions de paramètres pour l'architecture de classification seule). Il paraît tout à fait vraisemblable qu'une exploration plus exhaustive des architectures de GNN fournisse de bien meilleurs résultats. Cela se confirme par le fait qu'en comparant deux architectures (le PointNet++, architecture existante et éprouvée et le Attention-GNN qui est une personnalisation basée sur des opérateurs publiés plus récemment), nous avons bien montré que des gains de performances s'obtiennent grâce à de simples améliorations architecturales. Cela ouvre de vastes perspectives : il suffit pour cela de se référer à la littérature sur les CNNs et « d'emprunter » les techniques qui les ont rendus aussi performants : connexion résiduelle, normalisation de couche, diverses techniques de régularisation... Cet aspect pourra faire l'objet de futurs travaux.

7.1.2 Interprétabilité des modèles et Attention Concentrée

Au-delà des considérations générales sur les performances de modèles existants (CNNs ou ViT) présentées dans le chapitre 5, nos contributions propres -le pas adaptatif et son extension l'Attention Concentrée- méritent qu'on s'attarde sur quelques considérations. L'impact à la fois sur les performances (métriques brutes) et sur la qualité des cartes d'attributions

est largement démontré, tout particulièrement sur les images de fond d’œil. Certes, les attributions générées sont loin d’égaliser la précision des cartes obtenues avec un modèle de segmentation, mais elles s’en approchent plus que toutes les autres que nous avons comparées. Les résultats qualitatifs (figures 5.2 et 5.8) confortent le potentiel de l’approche, même si les gains quantitatifs par rapport aux méthodes existantes sont plus restreints en comparaison. Pour autant, la méthode de l’Attention Concentrée souffre de plusieurs limites qu’il est important de relever :

- Dans sa forme présentée ici, cette technique est relativement lente et coûteuse en ressources computationnelles. La raison est qu’en l’état, elle n’est pas parallélisable, autrement dit, l’algorithme d’Attention Concentrée ne peut s’appliquer que sur une image à la fois (et non par *mini-batch*)
- De plus, l’Attention Concentrée n’est utilisée que pour la génération a posteriori de cartes d’attributions, c’est-à-dire sur un modèle déjà entraîné. Or, l’échantillonnage conditionnel de jetons suivant une distribution d’importance conditionnelle pourrait a priori bénéficier à l’entraînement également.

Ce dernier point est particulièrement important : l’utilisation à l’inférence et pour la seule génération de cartes d’attributions est une restriction forte du potentiel de la méthode qui ne se justifie que pour des raisons purement techniques (notamment l’absence de parallélisation). En théorie, une architecture devrait bénéficier de cette technique aussi bien pour l’inférence que lors de l’apprentissage : l’introduction d’une forme de stochasticité dans le réseau est une technique éprouvée de régularisation conduisant à une meilleure généralisation ; en cela elle se compare au *Dropout* conditionnel. Par ailleurs, concentrer l’entraînement des ViT sur les jetons d’importance semble être pertinent pour accélérer l’apprentissage : plusieurs travaux s’y sont d’ailleurs depuis intéressés (Liang et al. [231], Yue et al. [232], Xu et al. [233]). Ces travaux, menés soit simultanément ou après les nôtres, n’utilisent cependant ni le pas adaptatif (permettant une approche multirésolution de la classification) ni l’échantillonnage aléatoire conditionnel comme régularisation.

CHAPITRE 8 CONCLUSION

Ainsi arrivons-nous à la conclusion de ce travail de doctorat, au terme duquel plusieurs contributions originales auront vu le jour, bâtissant -nous l'espérons- une armature solide par-dessus laquelle de futurs travaux pourront se greffer.

Pour reprendre une citation du poète Paul Valéry « *Il n'y a pas d'œuvre achevée, il n'y a que des œuvres abandonnées* ». Face à cette fatalité et afin d'éviter que l'inachèvement ne soit synonyme d'oubli, nous avons mis l'accent sur les limites de nos algorithmes, à la fois en matière de déploiement clinique qu'en discutant leurs débouchés pratiques ou théoriques. Cette analyse nous servira de support pour esquisser des améliorations possibles ou du moins les perspectives de recherche que nous jugeons pertinentes. Ces recommandations cloront cette thèse.

8.1 Contributions principales

À l'instar de notre champ d'étude, à cheval sur deux domaines, nous pouvons distinguer deux types de contributions : celles théoriques, appartenant à la recherche en informatique sur les réseaux de neurones et celles applicatives, plus proches du monde médical. Pour la première catégorie, nos contributions principales sont les suivantes :

1. La découverte que l'utilisation d'une attaque adversariale permet de faire une adaptation de style pour la segmentation sémantique (ce qui répond dans une certaine mesure à la définition de l'adaptation de domaine). Ce résultat est d'autant plus intrigant que l'attaque est réalisée à partir d'un modèle externe, en l'occurrence une simple sonde connectée dans l'encodeur de l'architecture de segmentation. Cette dernière n'est donc pas modifiée, mais il est cependant possible de lui faire adopter un style d'annotations au choix de l'utilisateur. Autre avantage de l'attaque adversariale : cette adaptation est humainement imperceptible et ne présente donc pas le risque de créer des structures « hallucinées » par le modèle à la différence des approches génératives. Ce résultat renseigne également sur le fonctionnement conjoint d'un encodeur et d'un décodeur au sein d'une architecture de segmentation : en effet, le décodeur n'est pas concerné par l'attaque (il n'y a pas de propagation des gradients en son sein, la sonde étant placée dans l'encodeur), mais malgré cela, il adoptera bien le style indiqué par l'encodeur.
2. La conception d'une représentation par graphe d'image basée sur les structures sémantiques qui la composent. En particulier, nous avons démontré qu'en utilisant deux

réseaux de CNNs classiques, un dédié à la segmentation, le second à la classification, nous n’observons pas de symbiose à l’inférence qui indiquerait la dépendance entre lésions segmentées et réseau de classification. Ce constat remet en question la pertinence de la segmentation dans un tel contexte. À l’inverse, la classification par graphe présente un biais inductif qui garantit cette dépendance. Outre la nouveauté de l’approche, celle-ci permet donc de garantir un lien causal entre structure segmentée et diagnostic, ce qui jalonne la voie vers une classification plus interprétable par l’utilisateur. Enfin, nous avons montré (en comparant PointNet++ et Attention-GNN) que les perspectives d’améliorations architecturales existent concernant les réseaux de graphes. Ceux-ci étant particulièrement légers et rapides à entraîner, nous espérons pouvoir significativement améliorer leurs performances à moindre coût.

3. Le second versant de ce travail doctoral porte sur l’interprétabilité des modèles de classification et plus spécifiquement, des architectures de type *Vision Transformer*. À cet égard, notre première contribution concerne la notion de pas adaptatif lors de l’étape de décomposition en jetons de l’image fournie au réseau. Différents pas (entraînant des séquences d’entrée plus ou moins longues) peuvent significativement altérer (en mieux ou en pire) les performances du réseau. À ce stade, le pas optimal peut donc être perçu comme un hyperparamètre supplémentaire à optimiser.
4. Dérivant de la notion de pas adaptatif, nous avons proposé un algorithme permettant le raffinement itératif de la carte d’attention, que nous avons appelé *Attention Concentrée*. L’algorithme introduit deux concepts fondamentaux : la notion d’attention multiéchelle (par l’utilisation d’un pas projectif de plus en plus petit) et la notion d’échantillonnage stochastique conditionnel. Cette dernière permet à la fois de modérer le coût computationnel de l’approche multiéchelle (qui reste ainsi de complexité constante), de concentrer le calcul de l’attribution locale sur les régions de l’image les plus pertinentes et d’agir à la manière d’une régularisation stochastique (ce dernier point restant cependant à démontrer formellement). L’Attention Concentrée permet la génération de cartes d’attributions de bien plus hautes résolutions que celles existantes dans la littérature sur les ViT.
5. Plus anecdotique, l’une des contributions les moins visibles de cette thèse concerne le développement de différentes plateformes qui ont permis la collaboration entre les différents membres des équipes médicales et notre laboratoire. Ce fut le cas pour la collecte d’annotations de segmentation, ainsi que pour le sondage sur la notion de qualité de l’interprétabilité. Ces outils posent en eux-mêmes des questions d’ingénierie et de recherche (par exemple sur les biais d’annotations que peuvent induire les interfaces des outils). La littérature ne prête pour l’instant qu’assez peu d’intérêt à

ces questions, mais nos outils (disponibles en accès libre) ont fait l'objet de nombreux échanges et mises à jour pour remplir leurs objectifs de recherche.

Enfin, au long du chapitre 6, nous avons également présenté brièvement diverses expériences qui se sont servies des algorithmes développés. Cela nous a permis d'aborder certaines des limites des algorithmes, reprises en détail dans notre discussion.

8.2 Recommandations

En conclusion, nous avons sélectionné les principaux éléments dont le développement nous paraît le plus prometteur, en leur adjoignant une série de recommandations pour de futurs travaux.

Classification par graphe rétinien La construction d'un graphe rétinien présente de nombreux avantages. Elle permet une représentation allégée de l'image, donc des modèles de classification extrêmement rapides, ce qui facilite leur déploiement sur des ordinateurs moins performants que ceux utilisés en recherche (notamment dans une volonté de décentralisation qui voudrait limiter le déplacement des données des patients hors de l'hôpital). Par ailleurs, la classification à partir de lésions ayant une sémantique clinique contraint le modèle à une forme de transparence du modèle appréciable pour les cliniciens. En revanche, les performances à ce jour sont loin d'égaler les modèles utilisant directement des images. Un des aspects qui devraient améliorer ces dernières est l'ajout dans le graphe d'autres biomarqueurs significatifs de l'image. En particulier, on pense à compléter le graphe avec les structures saines visibles dans le fond d'œil : arbre vasculaire, disque optique, macula... Ces éléments, intégrés au graphe rétinien, devraient a priori faciliter l'apprentissage par le GNN. Notons cependant que de nombreuses recherches restent à mener sur la bonne manière de construire ledit graphe, en particulier concernant les relations d'adjacences entre nœuds. Enfin, un autre axe qui mérite d'être mentionné est la notion d'interprétabilité appliquée aux GNNs. En effet, en utilisant des algorithmes similaires à ceux décrits pour les CNNs et ViTs, il serait possible d'identifier les nœuds du graphe ayant le plus contribué à la classification du modèle. Une telle information est extrêmement précieuse, non seulement pour comprendre la prise de décisions du réseau, mais également pour permettre de mieux comprendre l'évolution de la maladie diagnostiquée à partir de ses symptômes.

Attention Concentrée apprise Comme rappelé en discussion, l'algorithme d'Attention Concentrée a été développé dans le cadre de l'interprétabilité du modèle, c'est-à-dire pour la génération post-hoc de cartes d'attributions expliquant la prédiction du modèle. L'Attention

Concentrée ainsi développée n'est pas apprise et plus généralement, la notion d'échantillonnage conditionnel n'est pas utilisée ni lors de l'entraînement ni lors de l'inférence. Or, il existe des raisons crédibles de penser qu'elle pourrait bénéficier aux deux (actant comme une régularisation stochastique à l'apprentissage, permettant de limiter la taille de la matrice d'attention à calculer à un nombre fixe de jetons, etc.). De futurs travaux pourront se concentrer sur l'adaptation de l'Attention Concentrée la rendant fonctionnelle à l'apprentissage. À cette fin, nous avons d'ores et déjà développé une nouvelle version de l'algorithme, devenu parallélisable et optimisant ainsi très significativement son fonctionnement. L'entraînement d'un ViT de cette manière entraîne de substantiels gains en termes de performance. Cependant, une des principales difficultés reste la génération d'une carte d'échantillonnage conditionnelle pertinente durant l'apprentissage ; des recherches restent encore à mener sur cet aspect. On note simplement que des algorithmes comme l'*Attention Rollout* ne nécessitent pas de rétropropagation des gradients et pourraient donc facilement s'utiliser lors de la phase de propagation avant de l'entraînement.

DMLA La représentation par graphe est particulièrement propice pour représenter l'apparition de nouveaux vaisseaux dans la projection vasculaire de réseau choroïdien. En attachant à chaque arête des caractéristiques (par exemple, la tortuosité ou la longueur du vaisseau entre les deux nœuds aux extrémités de l'arête en question), le graphe devrait permettre la modélisation fine de l'état de l'arbre vasculaire. L'évolution temporelle peut alors se faire en faisant une comparaison longitudinale de graphes. Une métrique de proximité entre graphes n'est pas triviale à définir, mais une telle analyse est parfois faite à l'aide de GNN : cela correspond donc à une nouvelle application possible de l'architecture Attention-GNN développée dans le chapitre 4.

RÉFÉRENCES

- [1] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke et A. Rabinovich, “Going deeper with convolutions,” dans *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, p. 1–9. [En ligne]. Disponible : <http://ieeexplore.ieee.org/document/7298594/>
- [2] K. He, X. Zhang, S. Ren et J. Sun, “Deep Residual Learning for Image Recognition,” dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, p. 770–778. [En ligne]. Disponible : <http://ieeexplore.ieee.org/document/7780459/>
- [3] L.-C. Chen, G. Papandreou, F. Schroff et H. Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. [En ligne]. Disponible : <http://arxiv.org/abs/1706.05587>
- [4] G. Press. AI For Diabetic Retinopathy Screening Gets U.S. FDA Approval. Forbes. [En ligne]. Disponible : <https://www.forbes.com/sites/gilpress/2022/11/15/ai-for-diabetic-retinopathy-screening-gets-us-fda-approval/>
- [5] Eyenuk Announces FDA Clearance for EyeArt Autonomous AI System for Diabetic Retinopathy Screening. [En ligne]. Disponible : <https://www.businesswire.com/news/home/20200805005495/en/Eyenuk-Announces-FDA-Clearance-for-EyeArt-Autonomous-AI-System-for-Diabetic-Retinopathy-Screening>
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit et N. Houlsby, “An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale.” [En ligne]. Disponible : <https://openreview.net/forum?id=YicbFdNTTy>
- [7] C. E. Willoughby, D. Ponzin, S. Ferrari, A. Lobo, K. Landau et Y. Omid, “Anatomy and physiology of the human eye : Effects of mucopolysaccharidoses disease on structure and function – a review,” vol. 38, n°. s1, p. 2–11. [En ligne]. Disponible : <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1442-9071.2010.02363.x>
- [8] N. Panwar, P. Huang, J. Lee, P. A. Keane, T. S. Chuan, A. Richhariya, S. Teoh, T. H. Lim et R. Agrawal, “Fundus Photography in the 21st Century—A Review of Recent Technological Advances and Their Implications for Worldwide Healthcare,” vol. 22, n°. 3, p. 198–208. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4790203/>

- [9] Consortium of Neurology Clerkship Directors. [En ligne]. Disponible : <https://www.aan.com/residents-and-fellows/clerkship-and-course-director-resources/consortium-of-neurology-clerkship-directors/>
- [10] P. Thulasi, C. L. Fraser, V. Biousse, D. W. Wright, N. J. Newman et B. B. Bruce, “Nonmydriatic ocular fundus photography among headache patients in an emergency department,” vol. 80, n°. 5, p. 432–437. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3590046/>
- [11] V. Biousse, B. B. Bruce et N. J. Newman, “Ophthalmoscopy in the 21st century,” vol. 90, n°. 4, p. 167–175. [En ligne]. Disponible : <https://n.neurology.org/content/90/4/167>
- [12] D. P. Breen, C. W. Duncan, A. E. Pope, A. J. Gray et R. Al-Shahi Salman, “Emergency department evaluation of sudden, severe headache,” vol. 101, n°. 6, p. 435–443. [En ligne]. Disponible : <https://academic.oup.com/qjmed/article/101/6/435/1549287>
- [13] B. B. Bruce, P. Thulasi, C. L. Fraser, M. T. Keadey, A. Ward, K. L. Heilpern, D. W. Wright, N. J. Newman et V. Biousse, “Diagnostic accuracy and use of nonmydriatic ocular fundus photography by emergency physicians : Phase II of the FOTO-ED study,” vol. 62, n°. 1, p. 28–33.e1.
- [14] C. Lamirel, B. B. Bruce, D. W. Wright, K. P. Delaney, N. J. Newman et V. Biousse, “Quality of nonmydriatic digital fundus photography obtained by nurse practitioners in the emergency department : The FOTO-ED study,” vol. 119, n°. 3, p. 617–624.
- [15] B. B. Bruce, S. Bidot, R. Hage, L. C. Clough, C. Fajoles-Vasseneix, M. Melomed, M. T. Keadey, D. W. Wright, N. J. Newman et V. Biousse, “Fundus Photography vs. Ophthalmoscopy Outcomes in the Emergency Department (FOTO-ED) Phase III : Web-based, In-service Training of Emergency Providers,” vol. 42, n°. 5, p. 269–274. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6152496/>
- [16] Z. Yaqoob, J. Wu et C. Yang, “Spectral domain optical coherence tomography : A better OCT imaging strategy,” vol. 39, p. S6–S13. [En ligne]. Disponible : <https://www.future-science.com/doi/10.2144/000112090>
- [17] A. F. Fercher, K. Mengedocht et W. Werner, “Eye-length measurement by interferometry with partially coherent light,” vol. 13, n°. 3, p. 186–188. [En ligne]. Disponible : <https://www.osapublishing.org/ol/abstract.cfm?uri=ol-13-3-186>
- [18] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito et A. Et, “Optical coherence tomography,” vol. 254, n°. 5035, p. 1178–1181. [En ligne]. Disponible : <https://science.sciencemag.org/content/254/5035/1178>

- [19] A. F. Fercher, C. K. Hitzenberger, W. Drexler, G. Kamp et H. Sattmann, “In Vivo Optical Coherence Tomography,” vol. 116, n^o. 1, p. 113–114. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0002939414717623>
- [20] W. Drexler et J. G. Fujimoto, “State-of-the-art retinal optical coherence tomography,” vol. 27, n^o. 1, p. 45–88. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S1350946207000444>
- [21] A. Dubois, K. Grieve, G. Moneron, R. Lecaque, L. Vabre et C. Boccara, “Ultrahigh-resolution full-field optical coherence tomography,” vol. 43, n^o. 14, p. 2874–2883. [En ligne]. Disponible : <https://www.osapublishing.org/ao/abstract.cfm?uri=ao-43-14-2874>
- [22] A. Dubois, L. Vabre, A.-C. Boccara et E. Beaurepaire, “High-resolution full-field optical coherence tomography with a Linnik microscope,” vol. 41, n^o. 4, p. 805–812. [En ligne]. Disponible : <https://www.osapublishing.org/ao/abstract.cfm?uri=ao-41-4-805>
- [23] C. Hauger, M. Wörz et T. Hellmuth, “Interferometer for optical coherence tomography,” vol. 42, n^o. 19, p. 3896–3902. [En ligne]. Disponible : <https://www.osapublishing.org/ao/abstract.cfm?uri=ao-42-19-3896>
- [24] X. J. Wang, T. E. Milner et J. S. Nelson, “Characterization of fluid flow velocity by optical Doppler tomography,” vol. 20, n^o. 11, p. 1337–1339. [En ligne]. Disponible : <https://www.osapublishing.org/ol/abstract.cfm?uri=ol-20-11-1337>
- [25] R. A. Leitgeb, L. Schmetterer, C. K. Hitzenberger, A. F. Fercher, F. Berisha, M. Wojtkowski et T. Bajraszewski, “Real-time measurement of in vitro flow by Fourier-domain color Doppler optical coherence tomography,” vol. 29, n^o. 2, p. 171–173. [En ligne]. Disponible : <https://www.osapublishing.org/ol/abstract.cfm?uri=ol-29-2-171>
- [26] R. A. Leitgeb, R. M. Werkmeister, C. Blatter et L. Schmetterer, “Doppler Optical Coherence Tomography,” vol. 41, n^o. 100, p. 26–43. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4073226/>
- [27] D. Susan van, J. W. Beulens, S. Yvonne T. van der, D. E. Grobbee et B. Nealb, “The global burden of diabetes and its complications : An emerging pandemic,” vol. 17, p. s3–s8. [En ligne]. Disponible : <https://doi.org/10.1097/01.hjr.0000368191.86614.5a>
- [28] World Health Day 2016 : WHO calls for global action to halt rise in and improve care for people with diabetes. [En ligne]. Disponible : <https://www.who.int/news-room/detail/06-04-2016-world-health-day-2016-who-calls-for-global-action-to-halt-rise-in-and-improve-care-for-people-with-diabetes>

- [29] “Early Treatment Diabetic Retinopathy Study Design and Baseline Patient Characteristics : ETDRS Report Number 7,” vol. 98, p. 741–756. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0161642013380099>
- [30] P. Kolar, “Classification and Clinical Features of AMD.” [En ligne]. Disponible : <https://www.intechopen.com/books/age-related-macular-degeneration-etiology-diagnosis-and-management-a-glance-at-the-future/classification-and-clinical-features-of-amd>
- [31] W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng et T. Y. Wong, “Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040 : A systematic review and meta-analysis,” vol. 2, n^o. 2, p. e106–e116. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S2214109X13701451>
- [32] F. L. Ferris, C. P. Wilkinson, A. Bird, U. Chakravarthy, E. Chew, K. Csaky, S. R. Sadda et Beckman Initiative for Macular Research Classification Committee, “Clinical classification of age-related macular degeneration,” vol. 120, n^o. 4, p. 844–851.
- [33] Z. Lu, H. Pu, F. Wang, Z. Hu et L. Wang, “The expressive power of neural networks : A view from the width,” dans *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Curran Associates Inc., p. 6232–6240.
- [34] Y. Cai. Achieve the Minimum Width of Neural Networks for Universal Approximation. [En ligne]. Disponible : <http://arxiv.org/abs/2209.11395>
- [35] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie et L. Farhan, “Review of deep learning : Concepts, CNN architectures, challenges, applications, future directions,” vol. 8, n^o. 1, p. 53. [En ligne]. Disponible : <https://doi.org/10.1186/s40537-021-00444-8>
- [36] Y. Lecun, L. Bottou, Y. Bengio et P. Haffner, “Gradient-based learning applied to document recognition,” vol. 86, n^o. 11, p. 2278–2324.
- [37] A. Khan, A. Sohail, U. Zahoor et A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” vol. 53, n^o. 8, p. 5455–5516. [En ligne]. Disponible : <https://doi.org/10.1007/s10462-020-09825-6>
- [38] A. Krizhevsky, I. Sutskever et G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” dans *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc. [En ligne]. Disponible : <https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>
- [39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li et L. Fei-Fei, “ImageNet : A large-scale hierarchical image database,” dans *2009 IEEE Conference on Computer Vision and Pattern Recognition*, p. 248–255.

- [40] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever et R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” vol. abs/1207.0580. [En ligne]. Disponible : <http://arxiv.org/abs/1207.0580>
- [41] K. Simonyan et A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” dans *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio et Y. LeCun, édit. [En ligne]. Disponible : <http://arxiv.org/abs/1409.1556>
- [42] W. Luo, Y. Li, R. Urtasun et R. Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” dans *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS’16. Curran Associates Inc., p. 4905–4913.
- [43] X. Glorot et Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks.” [En ligne]. Disponible : <https://www.semanticscholar.org/paper/Understanding-the-difficulty-of-training-deep-Glorot-Bengio/b71ac1e9fb49420d13e084ac67254a0bbd40f83f>
- [44] K. He, X. Zhang, S. Ren et J. Sun, “Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification,” dans *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 1026–1034.
- [45] S. Ioffe et C. Szegedy, “Batch normalization : Accelerating deep network training by reducing internal covariate shift,” dans *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML’15. JMLR.org, p. 448–456.
- [46] D. Ulyanov, A. Vedaldi et V. Lempitsky, “Instance Normalization : The Missing Ingredient for Fast Stylization.” [En ligne]. Disponible : <https://www.semanticscholar.org/paper/Instance-Normalization%3A-The-Missing-Ingredient-for-Ulyanov-Vedaldi/63de0ad39d807f0c256f851428f211e8d5fcd3bb>
- [47] J. L. Ba, J. R. Kiros et G. E. Hinton. Layer Normalization. [En ligne]. Disponible : <http://arxiv.org/abs/1607.06450>
- [48] K. He et J. Sun, “Convolutional neural networks at constrained time cost,” dans *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 5353–5360.
- [49] R. K. Srivastava, K. Greff et J. Schmidhuber. Highway Networks. [En ligne]. Disponible : <http://arxiv.org/abs/1505.00387>
- [50] G. Huang, Z. Liu, L. vann der Maaten et K. Q. Weinberger, “Densely Connected Convolutional Networks,” p. 4700–4708. [En

- ligne]. Disponible : https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html
- [51] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang et X. Tang, “Residual Attention Network for Image Classification,” dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, p. 6450–6458. [En ligne]. Disponible : <https://ieeexplore.ieee.org/document/8100166/>
 - [52] J. Hu, L. Shen et G. Sun, “Squeeze-and-Excitation Networks,” dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 7132–7141.
 - [53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser et I. Polosukhin, “Attention is all you need,” dans *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, édit., vol. 30. Curran Associates, Inc. [En ligne]. Disponible : <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
 - [54] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng et S. Yan, “Tokens-to-token ViT : Training vision transformers from scratch on ImageNet,” dans *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 558–567.
 - [55] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei et Q. Tian, “Visformer : The vision-friendly transformer,” dans *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 589–598.
 - [56] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali et M. M. Fraz, “Vision Transformers in medical computer vision—A contemplative retrospection,” vol. 122, p. 106126. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S095219762300310X>
 - [57] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy et D. Merhof. Advances in Medical Image Analysis with Vision Transformers : A Comprehensive Review. [En ligne]. Disponible : <http://arxiv.org/abs/2301.03505>
 - [58] C. Farabet, C. Couprie, L. Najman et Y. LeCun, “Learning Hierarchical Features for Scene Labeling,” vol. 35, n°. 8, p. 1915–1929.
 - [59] J. Long, E. Shelhamer et T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” p. 3431–3440.
 - [60] O. Ronneberger, P. Fischer et T. Brox, “U-Net : Convolutional Networks for Biomedical Image Segmentation,” dans *Medical Image Computing and Computer-Assisted Inter-*

- vention – *MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells et A. F. Frangi, édit. Springer International Publishing, p. 234–241.
- [61] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi, M. C. Gilardi, S. Vitabile, G. Mauri, H. Nakayama et P. Cazzaniga, “USE-Net : Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets,” vol. 365, p. 31–43. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0925231219309245>
- [62] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker et D. Rueckert, “Attention U-Net : Learning Where to Look for the Pancreas.” [En ligne]. Disponible : <https://openreview.net/forum?id=Skft7cijM>
- [63] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh et J. Liang, “UNet++ : Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation,” vol. 39, n^o. 6, p. 1856–1867. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7357299/>
- [64] “Introduction,” dans *Numerical Optimization*, ser. Springer Series in Operations Research and Financial Engineering, J. Nocedal et S. J. Wright, édit. Springer, p. 1–9. [En ligne]. Disponible : https://doi.org/10.1007/978-0-387-40065-5_1
- [65] “Line Search Methods,” dans *Numerical Optimization*, ser. Springer Series in Operations Research and Financial Engineering, J. Nocedal et S. J. Wright, édit. Springer, p. 30–65. [En ligne]. Disponible : https://doi.org/10.1007/978-0-387-40065-5_3
- [66] A. Godichon-Baggioni, “Algorithmes stochastiques.” [En ligne]. Disponible : http://godichon.perso.math.cnrs.fr/Cours_algo_sto.pdf
- [67] I. Goodfellow, Y. Bengio et A. Courville, *Deep Learning*. MIT Press.
- [68] K. Kawaguchi, L. P. Kaelbling et Y. Bengio, “Generalization in Deep Learning.” [En ligne]. Disponible : <https://dspace.mit.edu/handle/1721.1/115274>
- [69] P. Bartlett, “The sample complexity of pattern classification with neural networks : The size of the weights is more important than the size of the network,” vol. 44, n^o. 2, p. 525–536.
- [70] N. Golowich, A. Rakhlin et O. Shamir, “Size-independent sample complexity of neural networks,” dans *Proceedings of the 31st Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet

- et P. Rigollet, édit., vol. 75. PMLR, p. 297–299. [En ligne]. Disponible : <https://proceedings.mlr.press/v75/golowich18a.html>
- [71] A. Blumer, A. Ehrenfeucht, D. Haussler et M. K. Warmuth, “Learnability and the Vapnik-Chervonenkis dimension,” vol. 36, n°. 4, p. 929–965. [En ligne]. Disponible : <https://dl.acm.org/doi/10.1145/76359.76371>
- [72] C. Zhang, S. Bengio, M. Hardt, B. Recht et O. Vinyals, “Understanding deep learning (still) requires rethinking generalization,” vol. 64, n°. 3, p. 107–115. [En ligne]. Disponible : <https://dl.acm.org/doi/10.1145/3446776>
- [73] T. Rostamizadeh, Mohri, *Foundations of Machine Learning, Second Edition*, 2^e éd., ser. MIT Press. [En ligne]. Disponible : <https://mitpress.ubli.sh.com/ebook/foundations-of-machine-learning--2-preview/7093/30>
- [74] D. Ramachandram et G. W. Taylor, “Deep Multimodal Learning : A Survey on Recent Advances and Trends,” vol. 34, n°. 6, p. 96–108.
- [75] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand et V. Lempitsky, “Domain-Adversarial Training of Neural Networks,” dans *Domain Adaptation in Computer Vision Applications*, G. Csurka, édit. Springer International Publishing, p. 189–209. [En ligne]. Disponible : http://link.springer.com/10.1007/978-3-319-58347-1_10
- [76] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros et T. Darrell, “CyCADA : Cycle-Consistent Adversarial Domain Adaptation,” dans *Proceedings of the 35th International Conference on Machine Learning*. PMLR, p. 1989–1998. [En ligne]. Disponible : <https://proceedings.mlr.press/v80/hoffman18a.html>
- [77] Y. Zhao, R. Wu et H. Dong, “Unpaired Image-to-Image Translation Using Adversarial Consistency Loss,” dans *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox et J.-M. Frahm, édit. Springer International Publishing, vol. 12354, p. 800–815. [En ligne]. Disponible : https://link.springer.com/10.1007/978-3-030-58545-7_46
- [78] Y. Zou, Z. Yu, B. V. K. Vijaya Kumar et J. Wang, “Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-training,” dans *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu et Y. Weiss, édit. Springer International Publishing, vol. 11207, p. 297–313. [En ligne]. Disponible : https://link.springer.com/10.1007/978-3-030-01219-9_18
- [79] Z. Zheng et Y. Yang, “Unsupervised Scene Adaptation with Memory Regularization in vivo,” dans *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, p. 1076–1082. [En ligne]. Disponible : <https://www.ijcai.org/proceedings/2020/150>

- [80] Y. Luo, G. Liu, Y. Guo et G. Yang, “Deep Neural Networks Learn Meta-Structures from Noisy Labels in Semantic Segmentation,” vol. 36, n^o. 2, p. 1908–1916. [En ligne]. Disponible : <https://ojs.aaai.org/index.php/AAAI/article/view/20085>
- [81] H. Song, M. Kim, D. Park, Y. Shin et J.-G. Lee, “Learning From Noisy Labels With Deep Neural Networks : A Survey,” p. 1–19.
- [82] M. Zhang, J. Gao, Z. Lyu, W. Zhao, Q. Wang, W. Ding, S. Wang, Z. Li et S. Cui, “Characterizing Label Errors : Confident Learning for Noisy-Labeled Image Segmentation,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, ser. Lecture Notes in Computer Science, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu et L. Joskowicz, édit. Springer International Publishing, p. 721–730.
- [83] Z. Xu, D. Lu, Y. Wang, J. Luo, J. Jayender, K. Ma, Y. Zheng et X. Li, “Noisy Labels are Treasure : Mean-Teacher-Assisted Confident Learning for Hepatic Vessel Segmentation,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, ser. Lecture Notes in Computer Science, p. u. Bruijne, Marleen, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng et C. Essert, édit. Springer International Publishing, p. 3–13.
- [84] R. Yi, Y. Huang, Q. Guan, M. Pu et R. Zhang, “Learning From Pixel-Level Label Noise : A New Perspective for Semi-Supervised Semantic Segmentation,” vol. 31, p. 623–635.
- [85] X.-J. Gui, W. Wang et Z.-H. Tian, “Towards Understanding Deep Learning from Noisy Labels with Small-Loss Criterion,” dans *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, p. 2469–2475. [En ligne]. Disponible : <https://www.ijcai.org/proceedings/2021/340>
- [86] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü et M. van Gerven, édit., *Explainable and Interpretable Models in Computer Vision and Machine Learning*, ser. The Springer Series on Challenges in Machine Learning. Springer International Publishing. [En ligne]. Disponible : <http://link.springer.com/10.1007/978-3-319-98131-4>
- [87] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto et D. Capurro, “A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images,” vol. 54, p. 203 :1–203 :40. [En ligne]. Disponible : <https://dl.acm.org/doi/10.1145/3522747>

- [88] E. Tjoa et C. Guan, “A Survey on Explainable Artificial Intelligence (XAI) : Toward Medical XAI,” vol. 32, n^o. 11, p. 4793–4813. [En ligne]. Disponible : <https://doi.org/10.1109/TNNLS.2020.3027314>
- [89] A. M. Groen, R. Kraan, S. F. Amirkhan, J. G. Daams et M. Maas, “A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology : Limited use of explainable AI?” vol. 157, p. 110592. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0720048X22004429>
- [90] A. Holzinger, G. Langs, H. Denk, K. Zatloukal et H. Müller, “Causability and explainability of artificial intelligence in medicine,” vol. 9, n^o. 4, p. e1312. [En ligne]. Disponible : <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1312>
- [91] D. W. Joyce, A. Kormilitzin, K. A. Smith et A. Cipriani, “Explainable artificial intelligence for mental health through transparency and interpretability for understandability,” vol. 6, n^o. 1, p. 1–7. [En ligne]. Disponible : <https://www.nature.com/articles/s41746-023-00751-9>
- [92] D. Jin, E. Sergeeva, W.-H. Weng, G. Chauhan et P. Szolovits, “Explainable deep learning in healthcare : A methodological survey from an attribution view,” vol. 14, n^o. 3, p. e1548. [En ligne]. Disponible : <https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.1548>
- [93] C. Patrício, J. C. Neves et L. F. Teixeira. Explainable Deep Learning Methods in Medical Imaging Diagnosis : A Survey. [En ligne]. Disponible : <http://arxiv.org/abs/2205.04766>
- [94] A. Singh, S. Sengupta et V. Lakshminarayanan, “Explainable Deep Learning Models in Medical Image Analysis,” vol. 6, n^o. 6, p. 52. [En ligne]. Disponible : <https://www.mdpi.com/2313-433X/6/6/52>
- [95] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi et J. Qadir, “Explainable, trustworthy, and ethical machine learning for healthcare : A survey,” vol. 149, p. 106043. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0010482522007569>
- [96] J. Petch, S. Di et W. Nelson, “Opening the Black Box : The Promise and Limitations of Explainable Machine Learning in Cardiology,” vol. 38, n^o. 2, p. 204–213. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0828282X21007030>
- [97] A. Barragán-Montero, A. Bibal, M. H. Dastarac, C. Draguet, G. Valdés, D. Nguyen, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, K. Souris, E. Sterpin et J. A. Lee, “Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability

- and data-model dependency,” vol. 67, n°. 11, p. 11TR01. [En ligne]. Disponible : <https://dx.doi.org/10.1088/1361-6560/ac678a>
- [98] Z. Salahuddin, H. C. Woodruff, A. Chatterjee et P. Lambin, “Transparency of deep neural networks for medical image analysis : A review of interpretability methods,” vol. 140, p. 105111. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0010482521009057>
- [99] S. Mohseni, N. Zarei et E. D. Ragan, “A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,” vol. 11, n°. 3-4, p. 24 :1–24 :45. [En ligne]. Disponible : <https://dl.acm.org/doi/10.1145/3387166>
- [100] G. Joshi, R. Walambe et K. Kotecha, “A Review on Explainability in Multimodal Deep Neural Nets,” vol. 9, p. 59 800–59 821.
- [101] N. Burkart et M. F. Huber, “A Survey on the Explainability of Supervised Machine Learning,” vol. 70, p. 245–317. [En ligne]. Disponible : <https://www.jair.org/index.php/jair/article/view/12228>
- [102] F. Doshi-Velez et B. Kim, “Considerations for Evaluation and Generalization in Interpretable Machine Learning,” dans *Explainable and Interpretable Models in Computer Vision and Machine Learning*, ser. The Springer Series on Challenges in Machine Learning, H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk et U. Güçlü, édit. Springer International Publishing, p. 3–17. [En ligne]. Disponible : https://doi.org/10.1007/978-3-319-98131-4_1
- [103] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan et R. Ranjan, “Explainable AI (XAI) : Core Ideas, Techniques, and Solutions,” vol. 55, n°. 9, p. 194 :1–194 :33. [En ligne]. Disponible : <https://doi.org/10.1145/3561048>
- [104] T. Speith, “A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods,” dans *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’22. Association for Computing Machinery, p. 2239–2250. [En ligne]. Disponible : <https://dl.acm.org/doi/10.1145/3531146.3534639>
- [105] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, p. u. Keulen, Maurice et C. Seifert, “From Anecdotal Evidence to Quantitative Evaluation Methods : A Systematic Review on Evaluating Explainable AI.” [En ligne]. Disponible : <https://dl.acm.org/doi/10.1145/3583558>
- [106] Z. C. Lipton, “The Mythos of Model Interpretability : In machine learning, the concept of interpretability is both important and slippery.” vol. 16, n°. 3, p. 31–57. [En ligne]. Disponible : <https://dl.acm.org/doi/10.1145/3236386.3241340>

- [107] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila et F. Herrera, “Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI,” vol. 58, n°. C, p. 82–115. [En ligne]. Disponible : <https://doi.org/10.1016/j.inffus.2019.12.012>
- [108] G. Montavon, W. Samek et K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” vol. 73, p. 1–15. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S1051200417302385>
- [109] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova et M. Prunotto, “Deep learning algorithm predicts diabetic retinopathy progression in individual patients,” vol. 2, n°. 1, p. 1–9. [En ligne]. Disponible : <https://www.nature.com/articles/s41746-019-0172-3>
- [110] D. Minh, H. X. Wang, Y. F. Li et T. N. Nguyen, “Explainable artificial intelligence : A comprehensive review,” vol. 55, n°. 5, p. 3503–3568. [En ligne]. Disponible : <https://doi.org/10.1007/s10462-021-10088-y>
- [111] M. Ghassemi, L. Oakden-Rayner et A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” vol. 3, n°. 11, p. e745–e750. [En ligne]. Disponible : [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(21\)00208-9/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext)
- [112] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz et H. A. Haenssle, “Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition,” vol. 155, n°. 10, p. 1135–1141.
- [113] M. D. Zeiler et R. Fergus, “Visualizing and Understanding Convolutional Networks,” dans *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele et T. Tuytelaars, édit. Springer International Publishing, p. 818–833.
- [114] K. Simonyan, A. Vedaldi et A. Zisserman, “Deep inside convolutional networks : Visualising image classification models and saliency maps,” dans *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, Y. Bengio et Y. LeCun, édit. [En ligne]. Disponible : <http://arxiv.org/abs/1312.6034>

- [115] J. Springenberg, A. Dosovitskiy, T. Brox et M. Riedmiller, “Striving for simplicity : The all convolutional net,” dans *ICLR (Workshop Track)*. [En ligne]. Disponible : <http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a>
- [116] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva et A. Torralba, “Learning Deep Features for Discriminative Localization.” IEEE Computer Society, p. 2921–2929. [En ligne]. Disponible : <https://www.computer.org/csdl/proceedings-article/cvpr/2016/8851c921/12OmNqIhFR6>
- [117] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh et D. Batra, “Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization,” dans *2017 IEEE International Conference on Computer Vision (ICCV)*, p. 618–626.
- [118] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller et W. Samek, “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” vol. 10, n°. 7, p. e0130140. [En ligne]. Disponible : <https://dx.plos.org/10.1371/journal.pone.0130140>
- [119] G. Montavon, S. Lapuschkin, A. Binder, W. Samek et K.-R. Müller, “Explaining nonlinear classification decisions with deep Taylor decomposition,” vol. 65, p. 211–222. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0031320316303582>
- [120] G. Montavon, A. Binder, S. Lapuschkin, W. Samek et K.-R. Müller, “Layer-Wise Relevance Propagation : An Overview,” dans *Explainable AI : Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen et K.-R. Müller, édit. Springer International Publishing, vol. 11700, p. 193–209. [En ligne]. Disponible : http://link.springer.com/10.1007/978-3-030-28954-6_10
- [121] H. Chefer, S. Gur et L. Wolf, “Transformer Interpretability Beyond Attention Visualization.” IEEE Computer Society, p. 782–791. [En ligne]. Disponible : <https://www.computer.org/csdl/proceedings-article/cvpr/2021/450900a782/1yeIsbbCMO4>
- [122] M. Sundararajan, A. Taly et Q. Yan, “Axiomatic attribution for deep networks,” dans *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML’17. JMLR.org, p. 3319–3328.
- [123] N. H. Shah, A. Milstein et P. Bagley, Steven C., “Making Machine Learning Models Clinically Useful,” vol. 322, n°. 14, p. 1351–1352. [En ligne]. Disponible : <https://doi.org/10.1001/jama.2019.10306>
- [124] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Büttner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, M. Wiesenfarth, E. Kavur, C. H. Sudre, M. Baumgartner, M. Eisenmann,

- D. Heckmann-Nötzel, A. T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. Blaschko, M. J. Cardoso, V. Cheplygina, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, p. u. Ginneken, Bram, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, H. Kenngott, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. I. Sánchez, S. Shetty, p. u. Smeden, Maarten, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. Van Calster, G. Varoquaux et P. F. Jäger, “Metrics reloaded : Pitfalls and recommendations for image analysis validation.” [En ligne]. Disponible : <https://ui.adsabs.harvard.edu/abs/2022arXiv220601653M>
- [125] A. Reinke, M. D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädsch, C. H. Sudre, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. Blaschko, F. Büttner, M. J. Cardoso, V. Cheplygina, J. Chen, E. Christodoulou, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, p. u. Ginneken, Bram, B. Glocker, P. Godau, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, H. Kenngott, J. Kleesiek, F. Kofler, T. Kooi, A. Kopp-Schneider, M. Kozubek, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, S. M. Rafelski, N. Rajpoot, M. Reyes, M. A. Riegler, N. Rieke, J. Saez-Rodriguez, C. I. Sánchez, S. Shetty, p. u. Smeden, Maarten, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. V. Calster, G. Varoquaux, M. Wiesenfarth, Z. R. Yaniv, P. F. Jäger et L. Maier-Hein. Understanding metric-related pitfalls in image analysis validation. ArXiv.
- [126] J. Cohen, “Weighted kappa : Nominal scale agreement provision for scaled disagreement or partial credit,” vol. 70, p. 213–220.
- [127] A. J. Conger, “Integration and generalization of kappas for multiple raters,” vol. 88, p. 322–328.
- [128] M. Shelley et K. Krippendorff, “Content analysis : An introduction to its methodology.” vol. 79, n°. 385, p. 240.
- [129] Y. Gal et Z. Ghahramani, “Dropout as a Bayesian Approximation : Representing Model Uncertainty in Deep Learning,” vol. 1, p. 2.

- [130] M. J. J. P. Van Grinsven, B. Van Ginneken, C. B. Hoyng, T. Theelen et C. I. Sánchez, “Fast Convolutional Neural Network Training Using Selective Data Sampling : Application to Hemorrhage Detection in Color Fundus Images,” vol. 35, n^o. 5, p. 1273–1284.
- [131] P. Chudzik, S. Majumdar, F. Caliva, B. Al-Diri et A. Hunter, “Exudate segmentation using fully convolutional neural networks and inception modules,” dans *Medical Imaging 2018 : Image Processing*, vol. 10574. SPIE, p. 785–792. [En ligne]. Disponible : <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10574/1057430/Exudate-segmentation-using-fully-convolutional-neural-networks-and-inception-modules/10.1117/12.2293549.full>
- [132] P. Chudzik, B. Al-Diri, F. Calivá, G. Ometto et A. Hunter, “Exudates Segmentation using Fully Convolutional Neural Network and Auxiliary Codebook,” dans *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, p. 770–773.
- [133] C. Payout, R. Duval et F. Cheriet, “A Multitask Learning Architecture for Simultaneous Segmentation of Bright and Red Lesions in Fundus Images,” dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. Lecture Notes in Computer Science, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López et G. Fichtinger, édit. Springer International Publishing, p. 101–108.
- [134] —, “A Novel Weakly Supervised Multitask Architecture for Retinal Lesions Segmentation on Fundus Images,” vol. 38, n^o. 10, p. 2434–2444.
- [135] Y. Xu, Z. Zhou, X. Li, N. Zhang, M. Zhang et P. Wei, “FFU-Net : Feature Fusion U-Net for Lesion Segmentation of Diabetic Retinopathy,” vol. 2021, p. 6644071. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7801055/>
- [136] S. Guo, T. Li, H. Kang, N. Li, Y. Zhang et K. Wang, “L-Seg : An end-to-end unified framework for multi-lesion segmentation of fundus images,” vol. 349, p. 52–63. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0925231219305430>
- [137] Z. Yan, X. Han, C. Wang, Y. Qiu, Z. Xiong et S. Cui, “Learning Mutually Local-Global U-Nets For High-Resolution Retinal Lesion Segmentation In Fundus Images,” dans *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, p. 597–600.
- [138] Y. Guo et Y. Peng, “CARNet : Cascade attentive RefineNet for multi-lesion segmentation of diabetic retinopathy images,” vol. 8, n^o. 2, p. 1681–1701. [En ligne]. Disponible : <https://doi.org/10.1007/s40747-021-00630-4>

- [139] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu et H. Kang, “Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening,” vol. 501, p. 511–522. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0020025519305377>
- [140] Y. Zhou, B. Wang, L. Huang, S. Cui et L. Shao, “A Benchmark for Studying Diabetic Retinopathy : Segmentation, Grading, and Transferability,” vol. 40, n^o. 3, p. 818–828.
- [141] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, T. Wu, J. Xiao, F. Wang, B. Yin, Y. Wang, G. Danala, L. He, Y. H. Choi, Y. C. Lee, S.-H. Jung, Z. Li, X. Sui, J. Wu, X. Li, T. Zhou, J. Toth, A. Baran, A. Kori, S. S. Chennamsetty, M. Safwan, V. Alex, X. Lyu, L. Cheng, Q. Chu, P. Li, X. Ji, S. Zhang, Y. Shen, L. Dai, O. Saha, R. Sathish, T. Melo, T. Araújo, B. Harangi, B. Sheng, R. Fang, D. Sheet, A. Hajdu, Y. Zheng, A. M. Mendonça, S. Zhang, A. Campilho, B. Zheng, D. Shen, L. Giancardo, G. Quéllec et F. Mériaudeau, “IDRiD : Diabetic Retinopathy – Segmentation and Grading Challenge,” vol. 59, p. 101561. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S1361841519301033>
- [142] Q. Wei, X. Li, W. Yu, X. Zhang, Y. Zhang, B. Hu, B. Mo, D. Gong, N. Chen, D. Ding et Y. Chen, “Learn to Segment Retinal Lesions and Beyond,” dans *2020 25th International Conference on Pattern Recognition (ICPR)*, p. 7403–7410.
- [143] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton et J.-C. Klein, “FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE : THE MESSIDOR DATABASE,” vol. 33, n^o. 3, p. 231–234. [En ligne]. Disponible : <https://www.ias-iss.org/ojs/IAS/article/view/1155>
- [144] S. Yamamoto et H. Yokouchi, “Automatic Recognition of Color Fundus Photographs,” dans *Digital Processing of Biomedical Images*, K. Preston et M. Onoe, édit. Springer US, p. 385–398. [En ligne]. Disponible : http://link.springer.com/10.1007/978-1-4684-0769-3_24
- [145] L. Guo, J.-J. Yang, L. Peng, J. Li et Q. Liang, “A computer-aided healthcare system for cataract classification and grading based on fundus image analysis,” vol. 69, p. 72–80. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0166361514001754>
- [146] R. Acharya U, C. K. Chua, E. Y. K. Ng, W. Yu et C. Chee, “Application of higher order spectra for the identification of diabetes retinopathy stages,” vol. 32, n^o. 6, p. 481–488.

- [147] U. R. Acharya, E. Y. K. Ng, J.-H. Tan, S. V. Sree et K.-H. Ng, “An Integrated Index for the Identification of Diabetic Retinopathy Stages Using Texture Parameters,” vol. 36, n°. 3, p. 2011–2020. [En ligne]. Disponible : <https://doi.org/10.1007/s10916-011-9663-8>
- [148] P. Chowriappa, S. Dua, U. Rajendra Acharya et M. Muthu Rama Krishnan, “Ensemble selection for feature-based classification of diabetic maculopathy images,” vol. 43, n°. 12, p. 2156–2162. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0010482513002813>
- [149] M. R. K. Mookiah, U. R. Acharya, H. Fujita, J. E. W. Koh, J. H. Tan, K. Noronha, S. V. Bhandary, C. K. Chua, C. M. Lim, A. Laude et L. Tong, “Local configuration pattern features for age-related macular degeneration characterization and classification,” vol. 63, p. 208–218. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0010482515002012>
- [150] J. Amin, M. Sharif et M. Yasmin, “A Review on Recent Developments for Detection of Diabetic Retinopathy,” vol. 2016, p. 6838976.
- [151] Y. Jeong, Y.-J. Hong et J.-H. Han, “Review of Machine Learning Applications Using Retinal Fundus Images,” vol. 12, n°. 1, p. 134.
- [152] A. Bilal, G. Sun et S. Mazhar, “Survey on recent developments in automatic detection of diabetic retinopathy,” vol. 44, n°. 3, p. 420–440.
- [153] S. Muchuchuti et S. Viriri, “Retinal disease detection using deep learning techniques : A comprehensive review,” vol. 9, p. 84.
- [154] K. Suriyasekeran, S. Santhanamahalingam et M. Duraisamy, “Algorithms for Diagnosis of Diabetic Retinopathy and Diabetic Macula Edema- A Review,” vol. 1307, p. 357–373.
- [155] S. N. Randive, R. K. Senapati et A. D. Rahulkar, “A review on computer-aided recent developments for automatic detection of diabetic retinopathy,” vol. 43, n°. 2, p. 87–99.
- [156] B. Goutam, M. F. Hashmi, Z. W. Geem et N. D. Bokde, “A Comprehensive Review of Deep Learning Strategies in Retinal Disease Diagnosis Using Fundus Images,” vol. 10, p. 57 796–57 823.
- [157] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega et D. R. Webster, “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” vol. 316, n°. 22, p. 2402–2410. [En ligne]. Disponible : <https://jamanetwork.com/journals/jama/fullarticle/2588763>

- [158] R. Gargeya et T. Leng, “Automated Identification of Diabetic Retinopathy Using Deep Learning,” vol. 124, n°. 7, p. 962–969. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0161642016317742>
- [159] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding et Y. Zheng, “Convolutional Neural Networks for Diabetic Retinopathy,” vol. 90, p. 200–205. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S1877050916311929>
- [160] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. S. Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee et T. Y. Wong, “Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes,” vol. 318, n°. 22, p. 2211–2223. [En ligne]. Disponible : <https://jamanetwork.com/journals/jama/fullarticle/2665775>
- [161] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang et M. He, “Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs,” vol. 125, n°. 8, p. 1199–1206. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0161642017335650>
- [162] J. Sahlsten, J. Jaskari, J. Kivinen, L. Turunen, E. Jaanio, K. Hietala et K. Kaski, “Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading,” vol. 9, n°. 1, p. 1–11. [En ligne]. Disponible : <https://www.nature.com/articles/s41598-019-47181-w>
- [163] F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm et B. H. F. Weber, “A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography,” vol. 125, n°. 9, p. 1410–1420. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0161642017330646>
- [164] M. D. Davis, R. E. Gangnon, L.-Y. Lee, L. D. Hubbard, B. E. K. Klein, R. Klein, F. L. Ferris, S. B. Bressler, R. C. Milton et Age-Related Eye Disease Study Group, “The Age-Related Eye Disease Study severity scale for age-related macular degeneration : AREDS Report No. 17,” vol. 123, n°. 11, p. 1484–1498.
- [165] S. Muchuchuti et S. Viriri, “Retinal Disease Detection Using Deep Learning Techniques : A Comprehensive Review,” vol. 9, n°. 4, p. 84. [En ligne]. Disponible : <https://www.mdpi.com/2313-433X/9/4/84>

- [166] J. Wu, R. Hu, Z. Xiao, J. Chen et J. Liu, “Vision Transformer-based recognition of diabetic retinopathy grade,” vol. 48, n°. 12, p. 7850–7863.
- [167] Z. Gu, Y. Li, Z. Wang, J. Kan, J. Shu et Q. Wang, “Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention,” vol. 2023, p. e1305583. [En ligne]. Disponible : <https://www.hindawi.com/journals/cin/2023/1305583/>
- [168] R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu et Y. Zhang, “Lesion-Aware Transformers for Diabetic Retinopathy Grading,” dans *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 10 933–10 942.
- [169] C. Adak, T. Karkera, S. Chattopadhyay et M. Saqib. Detecting Severity of Diabetic Retinopathy from Fundus Images using Ensembled Transformers. [En ligne]. Disponible : <http://arxiv.org/abs/2301.00973>
- [170] P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt et S. Farsiu, “Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images,” vol. 5, n°. 10, p. 3568–3577. [En ligne]. Disponible : <https://www.osapublishing.org/boe/abstract.cfm?uri=boe-5-10-3568>
- [171] G. Lemaître, M. Rastgoo, J. Massich, C. Y. Cheung, T. Y. Wong, E. Lamoureux, D. Milea, F. Mériaudeau et D. Sidibé. Classification of SD-OCT Volumes Using Local Binary Patterns : Experimental Validation for DME Detection. *Journal of Ophthalmology*. [En ligne]. Disponible : <https://www.hindawi.com/journals/joph/2016/3298606/>
- [172] A. Albarrak, F. Coenen, Y. Zheng *et al.*, “Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction,” dans *Proceedings of 2013 International Conference on Medical Image, Understanding and Analysis*, p. 59–64.
- [173] C. S. Lee, D. M. Baughman et A. Y. Lee, “Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images,” vol. 1, n°. 4, p. 322–327. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S2468653016301749>
- [174] S. P. K. Karri, D. Chakraborty et J. Chatterjee, “Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration,” vol. 8, n°. 2, p. 579–592.
- [175] R. A. Welikala, J. Dehmeshki, A. Hoppe, V. Tah, S. Mann, T. H. Williamson et S. A. Barman, “Automated detection of proliferative diabetic retinopathy using a

- modified line operator and dual classification,” vol. 114, n^o. 3, p. 247–261. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0169260714000650>
- [176] U. T. V. Nguyen, A. Bhuiyan, L. A. F. Park et K. Ramamohanarao, “An effective retinal blood vessel segmentation method using multi-scale line detection,” vol. 46, n^o. 3, p. 703–715. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S003132031200355X>
- [177] A. Christodoulidis, T. Hurtut, H. B. Tahar et F. Cheriet, “A multi-scale tensor voting approach for small retinal vessel segmentation in high resolution fundus images,” vol. 52, p. 28–43.
- [178] A. El Tanboly, M. Ismail, A. Shalaby, A. Switala, A. El-Baz, S. Schaal, G. Gimel’farb et M. El-Azab, “A computer-aided diagnostic system for detecting diabetic retinopathy in optical coherence tomography images,” vol. 44, n^o. 3, p. 914–23.
- [179] J. Novosel, K. A. Vermeer, p. u. Jong, J. H., Z. Wang et p. u. Vliet, L. J., “Joint Segmentation of Retinal Layers and Focal Lesions in 3-D OCT Data of Topologically Disrupted Retinas,” vol. 36, n^o. 6, p. 1276–1286.
- [180] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen et Z. Liu, “Attention to Lesion : Lesion-Aware Convolutional Neural Network for Retinal Optical Coherence Tomography Image Classification,” p. 1–1.
- [181] L. Huang, X. He, L. Fang, H. Rabbani et X. Chen, “Automatic Classification of Retinal Optical Coherence Tomography Images With Layer Guided Convolutional Neural Network,” vol. 26, n^o. 7, p. 1026–1030.
- [182] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger et N. Navab, “ReLayNet : Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks,” vol. 8, n^o. 8, p. 3627–3642. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5560830/>
- [183] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia et K. Zhang, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” vol. 172, n^o. 5, p. 1122–1131.e9. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0092867418301545>
- [184] M. C. Boucher, J. Qian, M. H. Brent, D. T. Wong, T. Sheidow, R. Duval, A. Kherani, R. Dookeran, D. Maberley, A. Samad, V. Chaudhary et Steering Committee for

- Tele-Ophthalmology Screening, Canadian Retina Research Network, “Evidence-based Canadian guidelines for tele-retina screening for diabetic retinopathy : Recommendations from the Canadian Retina Research Network (CR2N) Tele-Retina Steering Committee,” vol. 55, p. 14–24.
- [185] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez et P. Luo, “SegFormer : Simple and efficient design for semantic segmentation with transformers,” dans *Neural Information Processing Systems (NeurIPS)*.
 - [186] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin et M. Jorge Cardoso, “Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations,” vol. 2017, p. 240–248.
 - [187] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh et A. Talwalkar, “Hyperband : A Novel Bandit-Based Approach to Hyperparameter Optimization,” n°. 18.
 - [188] O. Ronneberger, P. Fischer et T. Brox, “U-Net : Convolutional Networks for Biomedical Image Segmentation,” dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells et A. F. Frangi, édit. Springer International Publishing, p. 234–241.
 - [189] S. Zagoruyko et N. Komodakis, “Wide Residual Networks,” dans *Proceedings of the British Machine Vision Conference 2016*. British Machine Vision Association, p. 87.1–87.12. [En ligne]. Disponible : <http://www.bmva.org/bmvc/2016/papers/paper087/index.html>
 - [190] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li et A. Smola, “ResNeSt : Split-Attention Networks,” dans *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, p. 2735–2745. [En ligne]. Disponible : <https://ieeexplore.ieee.org/document/9857221/>
 - [191] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez et P. Luo, “SegFormer : Simple and Efficient Design for Semantic Segmentation with Transformers.”
 - [192] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh et J. Liang, “UNet++ : A Nested U-Net Architecture for Medical Image Segmentation,” dans *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan et A. Madabhushi, édit. Springer International Publishing, p. 3–11.

- [193] S. Seferbekov, V. Iglovikov, A. Buslaev et A. Shvets, “Feature Pyramid Network for Multi-class Land Segmentation,” dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, p. 272–2723. [En ligne]. Disponible : <https://ieeexplore.ieee.org/document/8575511/>
- [194] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov et L.-C. Chen, “MobileNetV2 : Inverted Residuals and Linear Bottlenecks,” p. 4510–4520. [En ligne]. Disponible : http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html
- [195] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy et A. L. Yuille, “DeepLab : Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” vol. 40, n^o. 4, p. 834–848.
- [196] A. Garifullin, L. Lensu et H. Uusitalo, “Deep Bayesian baseline for segmenting diabetic retinopathy lesions : Advances and challenges,” vol. 136, p. 104725. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0010482521005199>
- [197] Z. Yan, X. Han, C. Wang, Y. Qiu, Z. Xiong et S. Cui, “Learning Mutually Local-Global U-Nets For High-Resolution Retinal Lesion Segmentation In Fundus Images,” dans *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, p. 597–600.
- [198] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui et L. Shao, “Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical Images,” dans *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, p. 2074–2083. [En ligne]. Disponible : <https://ieeexplore.ieee.org/document/8953911/>
- [199] G. Alain et Y. Bengio, “Understanding intermediate layers using linear classifier probes,” dans *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. [En ligne]. Disponible : <https://openreview.net/forum?id=HJ4-rAVtl>
- [200] A. Madry, A. Makelov, L. Schmidt, D. Tsipras et A. Vladu, “Towards deep learning models resistant to adversarial attacks,” dans *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. [En ligne]. Disponible : <https://openreview.net/forum?id=rJzIBfZAb>
- [201] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu et H. Kang, “Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening,” vol. 501, p. 511–522. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0020025519305377>

- [202] Q. Wei, X. Li, W. Yu, X. Zhang, Y. Zhang, B. Hu, B. Mo, D. Gong, N. Chen, D. Ding et Y. Chen, “Learn to Segment Retinal Lesions and Beyond.” IEEE Computer Society, p. 7403–7410. [En ligne]. Disponible : <https://www.computer.org/csdl/proceedings-article/icpr/2021/09412088/1tmj4hyd2w>
- [203] J. Cuadros et G. Bresnick, “EyePACS : An adaptable telemedicine system for diabetic retinopathy screening,” vol. 3, n^o. 3, p. 509–516.
- [204] APTOS 2019 Blindness Detection. [En ligne]. Disponible : <https://kaggle.com/c/aptos2019-blindness-detection>
- [205] M. C. Boucher, J. Qian, M. H. Brent, D. T. Wong, T. Sheidow, R. Duval, A. Kherani, R. Dookeran, D. Maberley, A. Samad et V. Chaudhary, “Evidence-based Canadian guidelines for tele-retina screening for diabetic retinopathy : Recommendations from the Canadian Retina Research Network (CR2N) Tele-Retina Steering Committee,” vol. 55, p. 14–24. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S0008418219312980>
- [206] C. R. Qi, L. Yi, H. Su et L. J. Guibas, “PointNet++ : Deep hierarchical feature learning on point sets in a metric space,” dans *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Curran Associates Inc., p. 5105–5114.
- [207] S. Brody, U. Alon et E. Yahav, “How attentive are graph attention networks?” dans *International Conference on Learning Representations*. [En ligne]. Disponible : <https://openreview.net/forum?id=F72ximsx7C1>
- [208] Y. Peskine, M.-C. Boucher et F. Cheriet, “An Interpretable Data-Driven Score for the Assessment of Fundus Images Quality,” dans *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, A. Campilho, F. Karray et Z. Wang, édit. Springer International Publishing, p. 325–331.
- [209] M. Bhaskaranand, C. Ramachandra, S. Bhat, J. Cuadros, M. G. Nittala, S. R. Sadda et K. Solanki, “The Value of Automated Diabetic Retinopathy Screening with the EyeArt System : A Study of More Than 100,000 Consecutive Encounters from People with Diabetes,” vol. 21, n^o. 11, p. 635–643. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6812728/>
- [210] C. J. Styles, “Introducing automated diabetic retinopathy systems : It’s not just about sensitivity and specificity,” vol. 33, n^o. 9, p. 1357–1358. [En ligne]. Disponible : <https://www.nature.com/articles/s41433-019-0535-7>
- [211] C. Matsoukas, J. F. Haslum, M. Söderberg et K. Smith, “Is it Time to Replace CNNs with Transformers for Medical Images?” dans *ICCV : Workshop on Computer*

- Vision for Automated Medical Diagnosis*, vol. abs/2108.09038. [En ligne]. Disponible : <https://arxiv.org/abs/2108.09038>
- [212] R. Wightman, “PyTorch Image Models,” Wightman. [En ligne]. Disponible : <https://github.com/rwightman/pytorch-image-models>
- [213] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens et Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2818–2826.
- [214] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin et A. A. Kalinin, “Albumentations : Fast and flexible image augmentations,” vol. 11, n^o. 2. [En ligne]. Disponible : <https://www.mdpi.com/2078-2489/11/2/125>
- [215] M. Melinščak, M. Radmilović, Z. Vataavuk et S. Lončarić, “Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation,” vol. 62, n^o. 3-4, p. 375–385. [En ligne]. Disponible : <https://doi.org/10.1080/00051144.2021.1973298>
- [216] K. He, X. Zhang, S. Ren et J. Sun, “Deep Residual Learning for Image Recognition,” dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770–778.
- [217] S. A. Kamran, S. Saha, A. S. Sabbir et A. Tavakkoli, “Optic-Net : A Novel Convolutional Neural Network for Diagnosis of Retinal Diseases from Optical Tomography Images,” dans *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, p. 964–971.
- [218] A. Singh, S. Sengupta, J. B. J, A. R. Mohammed, I. Faruq, V. Jayakumar, J. Zelek et V. Lakshminarayanan, “What is the Optimal Attribution Method for Explainable Ophthalmic Disease Classification?” dans *Ophthalmic Medical Image Analysis*. Springer, Cham, p. 21–31. [En ligne]. Disponible : https://link.springer.com/chapter/10.1007/978-3-030-63419-3_3
- [219] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan et O. Reblitz-Richardson, “Captum : A unified and generic model interpretability library for PyTorch.”
- [220] R. F. Spaide, G. J. Jaffe, D. Sarraf, K. B. Freund, S. R. Sadda, G. Staurenghi, N. K. Waheed, U. Chakravarthy, P. J. Rosenfeld, F. G. Holz, E. H. Souied, S. Y. Cohen, G. Querques, K. Ohno-Matsui, D. Boyer, A. Gaudric, B. Blodi, C. R. Bauman, X. Li, G. J. Coscas, A. Brucker, L. Singerman, P. Luthert, S. Schmitz-Valckenberg, U. Schmidt-Erfurth, H. E. Grossniklaus, D. J. Wilson, R. Guymer, L. A. Yannuzzi, E. Y. Chew, K. Csaky, J. M. Monés, D. Pauleikhoff, R. Tadayoni

- et J. Fujimoto, “Consensus Nomenclature for Reporting Neovascular Age-Related Macular Degeneration Data : Consensus on Neovascular Age-Related Macular Degeneration Nomenclature Study Group,” vol. 127, n^o. 5, p. 616–636. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0161642019322432>
- [221] M. Roy et M. Kaiser-Kupfer, “Second eye involvement in age-related macular degeneration : A four-year prospective study,” vol. 4 (Pt 6), p. 813–818.
- [222] R. E. Gangnon, K. E. Lee, B. E. K. Klein, S. K. Iyengar, T. A. Sivakumaran et R. Klein, “Severity of Age-Related Macular Degeneration in 1 Eye and the Incidence and Progression of Age-Related Macular Degeneration in the Fellow Eye,” vol. 133, n^o. 2, p. 125–132. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4326536/>
- [223] U. Schmidt-Erfurth, S. M. Waldstein, S. Klmscha, A. Sadeghipour, X. Hu, B. S. Gerendas, A. Osborne et H. Bogunović, “Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence,” vol. 59, n^o. 8, p. 3199–3208. [En ligne]. Disponible : <https://doi.org/10.1167/iovs.18-24106>
- [224] ASCRS : Prediction of imminent conversion to neovascular AMD using deep learning and 3D OCT volumes. Optometry Times. [En ligne]. Disponible : <https://www.optometrytimes.com/view/ascrs-prediction-of-imminent-conversion-to-neovascular-amd-using-deep-learning-and-3d-oct-volumes>
- [225] J. Yim, R. Chopra, T. Spitz, J. Winkens, A. Obika, C. Kelly, H. Askham, M. Lukic, J. Huemer, K. Fasler, G. Moraes, C. Meyer, M. Wilson, J. Dixon, C. Hughes, G. Rees, P. T. Khaw, A. Karthikesalingam, D. King, D. Hassabis, M. Suleyman, T. Back, J. R. Ledsam, P. A. Keane et J. De Fauw, “Predicting conversion to wet age-related macular degeneration using deep learning,” vol. 26, n^o. 6, p. 892–899. [En ligne]. Disponible : <https://www.nature.com/articles/s41591-020-0867-7>
- [226] L. Najman et M. Schmitt, “Watershed of a continuous function,” vol. 38, n^o. 1, p. 99–112. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S0165168494900590>
- [227] S. Panchal, A. Naik, M. Kokare, S. Pachade, R. Naigaonkar, P. Phadnis et A. Bhange, “Retinal Fundus Multi-Disease Image Dataset (RFMiD) 2.0 : A Dataset of Frequently and Rarely Identified Diseases,” vol. 8, n^o. 2, p. 29. [En ligne]. Disponible : <https://www.mdpi.com/2306-5729/8/2/29>
- [228] H. Daumé III, “Frustratingly Easy Domain Adaptation,” dans *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

- Association for Computational Linguistics, p. 256–263. [En ligne]. Disponible : <https://aclanthology.org/P07-1033>
- [229] S. A. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. H. Maier-Hein, g.-i. Eslami, SM, D. J. Rezende et O. Ronneberger, “A probabilistic U-net for segmentation of ambiguous images.”
- [230] J. Choi, T. Kim et C. Kim, “Self-Ensembling With GAN-Based Data Augmentation for Domain Adaptation in Semantic Segmentation,” dans *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, p. 6829–6839. [En ligne]. Disponible : <https://ieeexplore.ieee.org/document/9008277/>
- [231] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang et P. Xie, “Not All Patches are What You Need : Expediting Vision Transformers via Token Reorganizations.”
- [232] X. Yue, S. Sun, Z. Kuang, M. Wei, P. H. S. Torr, W. Zhang et D. Lin, “Vision Transformer With Progressive Sampling,” p. 387–396.
- [233] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu et X. Sun, “Evo-vit : Slow-fast token evolution for dynamic vision transformer,” dans *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, n°. 3, p. 2964–2972.
- [234] M. D. Abràmoff, M. K. Garvin et M. Sonka, “Retinal Imaging and Image Analysis,” vol. 3, p. 169–208. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3131209/>
- [235] J. Xu, O. Chutatape, C. Zheng et P. C. T. Kuan, “Three dimensional optic disc visualisation from stereo images via dual registration and ocular media optical correction,” vol. 90, n°. 2, p. 181–185. [En ligne]. Disponible : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1860159/>
- [236] M. Yung, M. A. Klufas et D. Sarraf, “Clinical applications of fundus autofluorescence in retinal disease,” vol. 2, n°. 1, p. 12. [En ligne]. Disponible : <https://doi.org/10.1186/s40942-016-0035-x>
- [237] Z. Zhang, J. Liu, F. Yin, B.-H. Lee, D. W. K. Wong et K. R. Sung, “ACHIKO-K : Database of fundus images from glaucoma patients,” dans *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, p. 228–231.
- [238] A. Hoover, V. Kouznetsova et M. Goldbaum, “Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response,” vol. 19, n°. 3, p. 203–210.
- [239] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever et p. u. Ginneken, B., “Ridge-based vessel segmentation in color images of the retina,” vol. 23, n°. 4, p. 501–509.

- [240] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, E. Fletcher et L. Kennedy, "Optic Nerve Head Segmentation," vol. 23, n^o. 2, p. 256–264. [En ligne]. Disponible : <http://ieeexplore.ieee.org/document/1263614/>
- [241] R. Zwiggelaar, édit., *Medical Image Understanding and Analysis 2007 : University of Wales Aberystwyth, 17 - 18th July*. BMVA.
- [242] T. Kauppi, V. Kalesnykienne, J.-K. Kamareinen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, J. Pietilä, H. Kälviäinen et H. Uusitalo, "DIARETDB1 diabetic retinopathy database and evaluation protocol," vol. 2007, p. 61.
- [243] E. J. Carmona, M. Rincón, J. García-Feijoó et J. M. Martínez-de-la Casa, "Identification of the optic nerve head with genetic algorithms," vol. 43, n^o. 3, p. 243–259.
- [244] M. Niemeijer, p.-u. Ginneken, Bram, M. J. Cree, A. Mizutani, G. Quellec, C. I. Sanchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, X. Wu, G. Cazuguel, J. You, A. Mayo, Q. Li, Y. Hatanaka, g.-i. Cochener, Béatrice, C. Roux, F. Karray, M. Garcia, H. Fujita et M. D. Abramoff, "Retinopathy Online Challenge : Automatic Detection of Microaneurysms in Digital Color Fundus Photographs," vol. 29, n^o. 1, p. 185–195.
- [245] L. Tang, M. K. Garvin, K. Lee, W. L. Alward, Y. H. Kwon et M. D. Abramoff, "Robust Multiscale Stereo Matching from Fundus Images with Radiometric Differences," vol. 33, n^o. 11, p. 2245–2258.
- [246] M. Niemeijer, X. Xu, A. V. Dumitrescu, P. Gupta, p.-u. Ginneken, Bram, J. C. Folk et M. D. Abramoff, "Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs," vol. 30, n^o. 11, p. 1941–1950.
- [247] F. Fumero, S. Alayon, J. L. Sanchez, J. Sigut et M. Gonzalez-Hernandez, "RIM-ONE : An open retinal image database for optic nerve evaluation," dans *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, p. 1–6.
- [248] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J. C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ et A. Chabouis, "TeleOphta : Machine learning and image processing methods for teleophthalmology," vol. 34, n^o. 2, p. 196–203. [En ligne]. Disponible : <http://www.sciencedirect.com/science/article/pii/S1959031813000237>
- [249] T. Kohler, A. Budai, M. F. Kraus, J. Odstrcilik, G. Michelson et J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," dans *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, p. 95–100. [En ligne]. Disponible : <http://ieeexplore.ieee.org/document/6627771/>

- [250] J. Sivaswamy, A. Chakravarty, G. D. Joshi et T. A. Syed, “A Comprehensive Retinal Image Dataset for the Assessment of Glaucoma from the Optic Nerve Head Analysis.”
- [251] S. Joshi et P. T. Karule, “Review of Preprocessing Techniques for Fundus Image Analysis,” vol. 60, n°. 3, p. 593–612. [En ligne]. Disponible : http://www.iieta.org/journals/ama_b/paper/10.18280/ama_b.600306
- [252] N. Mukherjee et S. Sengupta, “In Search for the Optimal Preprocessing Technique for Deep Learning Based Diabetic Retinopathy Stage Classification from Fundus Images.” [En ligne]. Disponible : <https://www.researchsquare.com/article/rs-654484/v1>
- [253] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet et J. M. P. Langlois, “Red Lesion Detection Using Dynamic Shape Features for Diabetic Retinopathy Screening,” vol. 35, n°. 4, p. 1116–1126.
- [254] A. Hoover et M. Goldbaum, “Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels,” vol. 22, n°. 8, p. 951–958.
- [255] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith *et al.*, “Model soups : Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time,” dans *International Conference on Machine Learning*. PMLR, p. 23 965–23 998.
- [256] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira et Y. Wu, “Contrastive Learning for Label Efficient Semantic Segmentation,” dans *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, p. 10 603–10 613. [En ligne]. Disponible : <https://ieeexplore.ieee.org/document/9710192/>

ANNEXE A MODALITÉS D'IMAGERIE DE LA RÉTINE

En complément de la section 2.2, nous proposons une brève introduction des autres modalités existantes permettant d'obtenir une acquisition de la rétine. La structuration des informations fournies dans cette section suit celle de Abramoff et al. [234], qui est une revue de littérature de 2010 portant sur l'imagerie rétinienne et les traitements d'image associés.

Le fond d'œil hyperspectral (ou multispectral) propose d'étendre la sensibilité du capteur à d'autres longueurs d'ondes du visible (autres que le RVB) allant typiquement de 495 nm jusqu'à 570 nm. Pour contrôler précisément la longueur d'onde que l'on souhaite mesurer, un laser accordable est employé. L'acquisition hyperspectrale ne résulte pas en une seule image mais un ensemble d'images (chacune correspondant à une certaine longueur) d'onde. Généralement, la résolution spectrale est de l'ordre de 5 nm ; la principale difficulté étant de réaliser un nombre d'acquisitions élevé en un temps minimal (pour limiter les perturbations liées au mouvement du patient). A ce stade, les débouchés cliniques de l'imagerie hyperspectrale sont en phase de recherche.

L'acquisition stéréographique Dans une certaine mesure, l'acquisition de fond d'œil permet également de reconstruire une approximation tri-dimensionnelle de la surface par usage de la stéréographie. Cette approche repose sur l'usage de deux photographies, pris à des angles différents, pour estimer la profondeur locale. Différentes étapes sont mises en jeu pour cela, présentées dans les travaux de Xu et al. [235] : en premier lieu, la caméra est calibrée afin d'en extraire les paramètres optiques. Les deux acquisitions, faites selon deux inclinaisons de la caméra, sont prises de la rétine. Une reconstruction 3D est ensuite effectuée entre les deux photographies, permettant ainsi d'extraire l'information de profondeur.

La fundoscopie par autofluorescence L'imagerie de Fundoscopie par Autofluorescence (FAF) est une approche non-invasive permettant d'imager des structures spécifiques de la rétine. A la différence de l'imagerie de fond d'œil classique, elle ne mesure cependant pas une réflexion lumineuse mais une émission lumineuse auto-émise par la structure que l'on cible. Son principe est d'exciter un fluorophore par une onde lumineuse excitatrice (communément de longueur d'onde correspondant au bleu). Cet apport d'énergie à une longueur d'onde spécifique élève le niveau d'énergie des électrons composant le fluorophore. En se désexcitant

(c'est à dire en retournant à leur niveau d'énergie initial), les électrons émettent un rayonnement lumineux qui sera mesuré. Le fluorophore visé -la lipofuscine- étant essentiellement présent dans la RPE, cette technique est particulièrement adaptée pour le diagnostic de la DMLA. La FAF n'est cependant pas limitée à ce diagnostic. En parcourant d'autres longueurs d'ondes excitatrices, correspondant à d'autres fluorophores spécifiques, faisant ainsi ressortir des marqueurs symptomatiques d'autres maladies. Nous recommandons le travail de Yung et al. [236] pour des illustrations comparées des images obtenues par fond d'œil classique et par autofluorescence.

ANNEXE B BASES DE DONNÉES DE FOND D'ŒIL PUBLIQUEMENT ACCESSIBLES

Nom	Référence	Nombre d'images	Type de vérité terrain.
ACHIKO-K	[237]	258	Masque de segmentation de symptômes propres au glaucome
STARE	[238]	402	Diagnostic de maladie + divers masques sur quelques images
DRIVE	[239]	40	Masque sémantique du réseau vasculaire
ONHSD	[240]	99	Masque du disque optique
DIARETDB0	[241]	130	Texte indiquant la présence de lésions par image
DIARETDB1		89	Annotations des régions de l'image contenant des lésions
DRIONS-DB	[243]	110	Masque du disque optique
ROC	[244]	100	Coordonnées des micro-anévrismes
INSPIRE-stereo	[245]	30 (stéréo)	Profondeur obtenue par OCT
INSPIRE-AVR	[246]	40	Masque du réseau vasculaire et Ratio artère-veine (<i>Artery-vein ratio</i>) (AVR)
RIM-ONE	[247]	159	Masque du disque optique
E-OPHTA (EX)	[248]	82	Masque des exsudats durs
E-OPHTA (MA)		381	Masque des micro-anévrismes
HRF - segmentation	[249]	45	Masque du réseau vasculaire

HRF - quality		18 (mauvaise qualité)	18 images reprises (bonne qualité)
MESSIDOR-2	[143]	874×2 (paire D/G)	Grade de la RD et risque de OMD.
Drishti-GS1	[250]	101	Masque du disque optique et de l'excavation papillaire
EyePACS - Kaggle	EyePACS (2015)	88702	Grade de la RD.
Aptos - Kaggle	Aptos (2019)	4000	Grade de la RD.
IDRID	[141]	81	Segmentation lésions
		516	Grade de la RD
		516	Localisation OD/Macula
DDR	[139]	757	Segmentation lésions
		13673	Grade de la RD
RETINAL-LESIONS	[142]	1593	Segmentation des lésions + gradation RD.
FGADR	[140]	1842	Segmentation lésions
		1000	Grade de la RD

ANNEXE C RÉSULTATS COMPLÉMENTAIRES SUR LA SEGMENTATION DU FOND D'ŒIL

C.1 Homogénéisation de l'apparence

La notion d'apparence d'une image n'a pas de définition établie car elle repose en partie sur une appréciation subjective de concepts tels que le contraste, l'exposition, le contenu textural, fréquentiel... En pratique, il est néanmoins relativement aisé de reconnaître sans ambiguïté la base de données d'appartenance d'une image. Or, à des fins de généralisation, cette propriété n'est pas souhaitable : idéalement l'origine de la donnée devrait être décorrélée de l'analyse qui en est faite. Pour garantir que les algorithmes se comportent de la manière la plus prédictible possible, un certain nombre d'articles propose d'uniformiser les images via un prétraitement. On peut noter que cet argument n'est souvent que secondaire, et bien souvent, c'est la capacité du prétraitement à faire ressortir les structures importantes ou à corriger des défauts d'acquisition (illumination non-homogène notamment) qui est mise en avant [251,252]. Quoiqu'il en soit, il n'existe aucun consensus clair sur l'intérêt de ces prétraitements dans un cadre général, c'est à dire indépendamment de la tâche et du modèle de classification considérés. Les travaux les plus significatifs présentés dans la revue de littérature sur la segmentation (section 3.5.1) et la classification (section 3.5.3) font souvent l'économie de cette étape. Pourtant, vis-à-vis de notre problématique sur la généralisation inter-bases, l'hypothèse d'une amélioration des résultats via l'homogénéisation de l'apparence des images mérite d'être étudiée. Nous adoptons un protocole similaire à celui de Séoud et al. [253], tout en y ajoutant une étape d'égalisation d'histogramme. Différentes constantes (obtenues relativement à la résolution de l'image) sont utilisées dans le protocole de prétraitement et sont définies tel qu'il suit :

- D , le diamètre de la région d'intérêt. Comme la majorité des images à notre disposition sont prises avec un angle d'acquisition (Field of View (FOV)) constant de 45° , D permet la paramétrisation des filtres de manière invariante à la résolution des images.
- $d_1 = D/10$, l'estimation du rayon moyen du disque optique.
- $d_2 = D/360$, la plus petite taille de micro-anévrysmes décelable.
- $d_3 = D/28$, la plus grande taille d'hémorragies attendue.

Correction de l'illumination Hoover et al. [254] suggèrent de compenser l'effet de vignettage (bordures assombries aux extrémités de la région d'intérêt) par une correction de la moyenne locale vers la moyenne globale de l'image. Originellement, la moyenne locale était

estimée par une convolution par un filtre moyen de taille d_1 . Nous la remplaçons par un filtre médian, moins sensible aux effets de bords et permettant de mieux préserver les contours de l'image. Le filtre médian, noté h_{M1} est choisi de taille d_1

$$I_e = I + \mu - h_{M1}(I) \quad (\text{C.1})$$

μ représente la valeur moyenne du canal considéré. Cette égalisation est réalisée sur les trois canaux séparément.

Débruitage Afin de réduire le bruit gaussien, un petit filtre moyennneur h_{M2} de taille d_2 est appliqué à l'image I_e .

$$I_{dn} = I_e * h_{M2} \quad (\text{C.2})$$

Homogénéisation adaptative des contrastes Les zones de faible contraste sont identifiées par l'évaluation locale d'écart-type I_{std} dans une région de taille d_2 . Une valeur faible de cette carte d'écart-type indique soit une zone homogène, soit une zone de faible contraste. Pour ces dernières, les détails sont renforcés par l'utilisation d'un filtre passe-haut.

Standardisation des couleurs Chacun des canaux est standardisé autour d'une même valeur moyenne μ , ce qui a pour effet de centrer la distribution des valeurs de pixels. De plus, les valeurs de cet histogramme dépassant de $\pm 3\sigma$ la valeur moyenne sont tronquées.

Égalisation d'histogramme adaptative à contraste limité L'égalisation d'histogramme est une technique classique de renforcement des contrastes. Dans sa version classique, il s'agit de trouver la transformation qui mappe l'histogramme de l'image vers une approximation de distribution uniforme, autrement dit, qui vienne étaler l'histogramme sur l'ensemble des valeurs possibles, augmentant ainsi le contraste de l'image. Cette opération est facile à implémenter en pratique en faisant intervenir les propriétés de l'inverse d'une fonction de distribution.

Dans sa version adaptative, cette égalisation se fait localement dans une fenêtre donnée. Mais si cette fenêtre couvre une zone homogène de l'image, l'égalisation d'histogramme adaptative produit généralement un résultat bruité (le bruit d'acquisition étant fortement amplifié). L'algorithme Contrast Limited Adaptive Histogram Equalization (CLAHE) corrige cet effet en limitant l'amplification du contraste par un effet seuillage sur les pics d'histogramme. Nous l'appliquons sur le canal de luminance de l'image obtenue après sa décomposition LAB. Les différentes étapes sont illustrées sur la figure C.1. La figure C.2 permet d'en apprécier

l'effet sur chacune des bases de données.

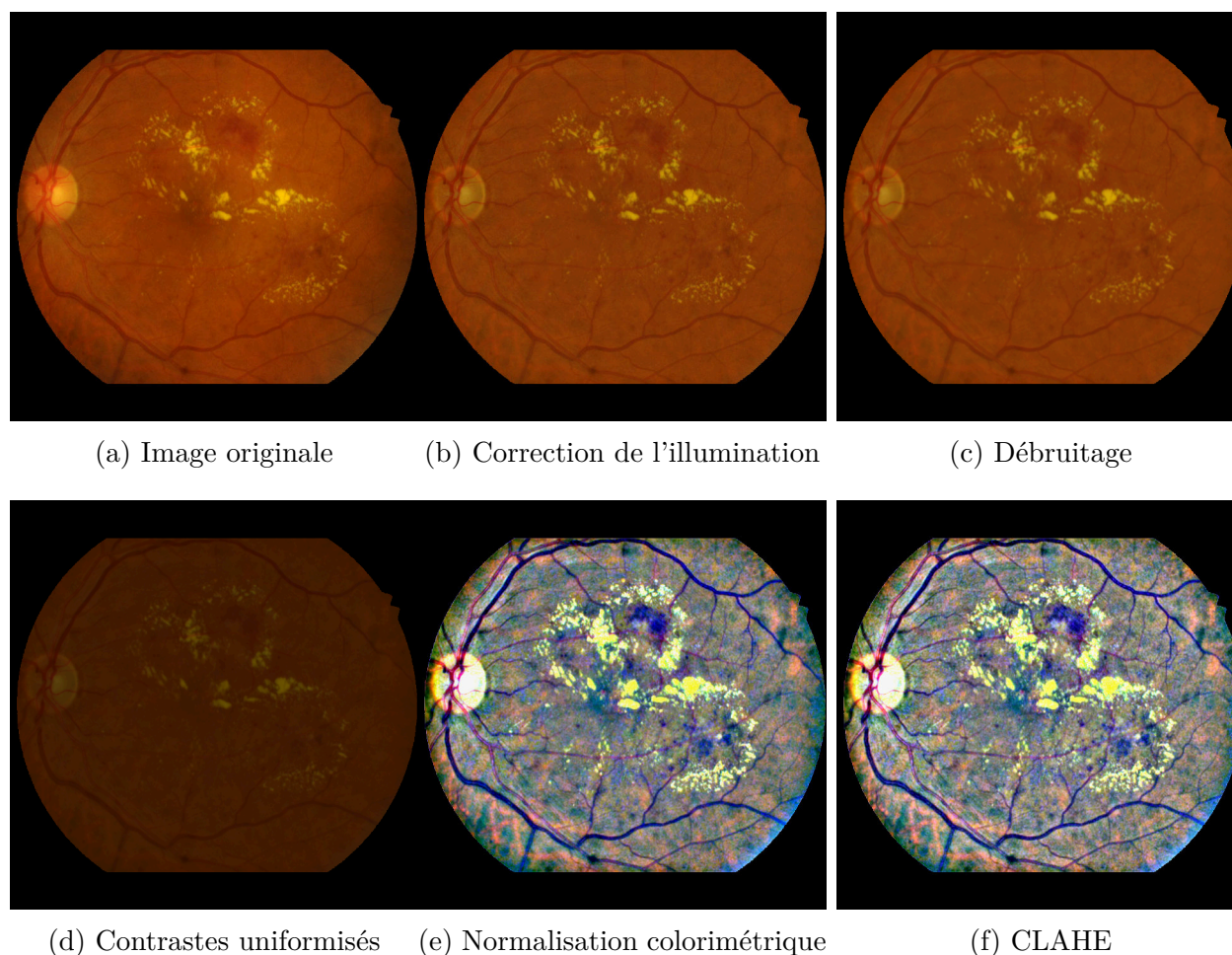


FIGURE C.1 Les différentes étapes du prétraitement appliqué aux images de fond d'œil.

C.1.1 Impact expérimental

Le prétraitement appliqué vise à standardiser les images des différentes bases de données. S'il n'existe pas de métrique pouvant en apprécier l'efficacité, on en peut en mesurer l'effet de manière indirecte. La figure C.3 illustre les histogrammes des images pour chacune des bases de données. Même s'il est possible de distinguer des similarités dans les distributions avant prétraitement, celles-ci deviennent évidentes après application de celui-ci, les histogrammes tendent à s'aligner très fortement. À noter qu'il s'agit d'une mesure indirecte de la similarité entre bases, les histogrammes ayant statistiquement tendance à converger vers une distribution normale qui ne peut pas réellement révéler les disparités individuelles entre images.

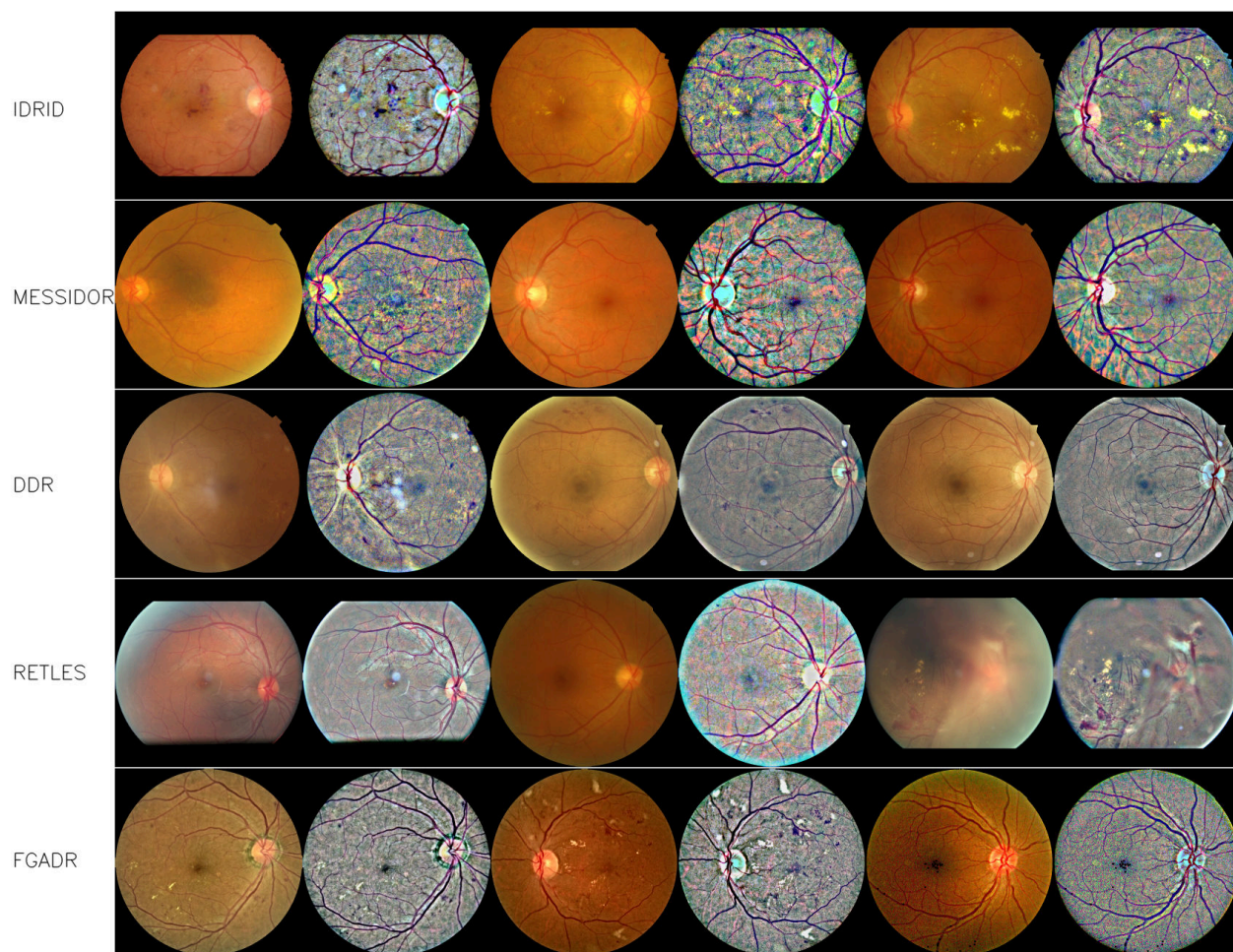


FIGURE C.2 Chaque ligne correspond à une base de données. Le prétraitement proposé vise à homogénéiser les images pour limiter la variabilité inter et intra bases. Il a aussi pour effet d'uniformiser la luminosité et le contraste au sein de l'image.

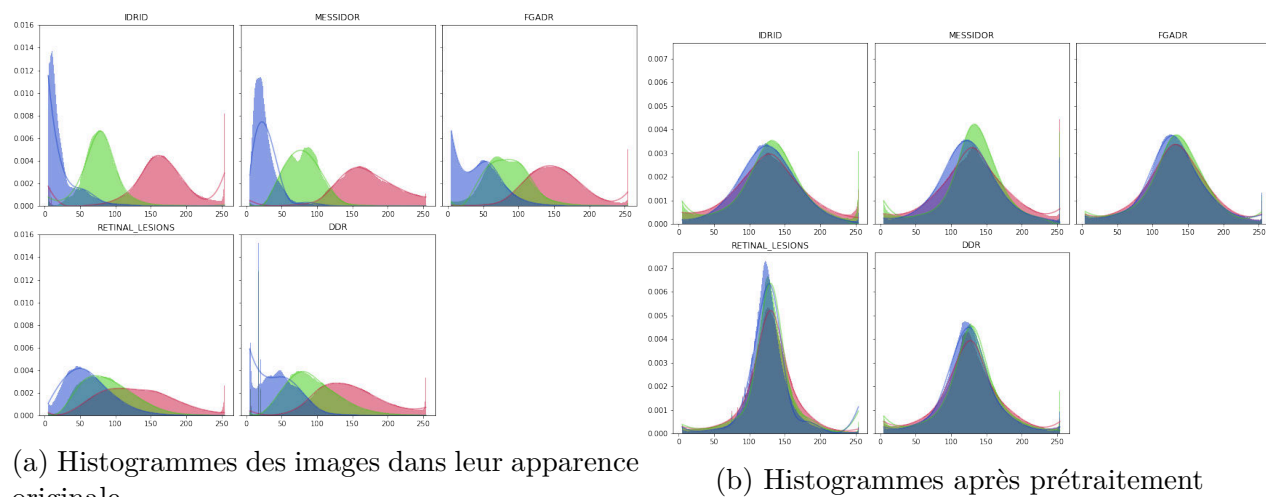


FIGURE C.3 Histogramme des images de chaque base de données, avant et après prétraitement.

Il est en revanche plus pertinent de comparer les performances d'un modèle entraîné avec ou sans prétraitement. Là encore, nous faisons le choix de comparer les modèles en fonction de leurs données d'entraînement, en différenciant les deux scénarios. Les graphiques rapportés sur la figure C.4 fournit le détail de ces résultats (*attention, à la différence des graphiques de la section 4.3.2, les abscisses représentent les bases d'évaluation*). Il y apparaît clairement que le prétraitement proposé à tendance, globalement, à diminuer les performances du modèle, bien que relativement marginalement. Il existe cependant une exception notable à ce constat : le modèle entraîné avec FGADR voit ses performances considérablement améliorées sur l'ensemble des bases d'évaluation avec le prétraitement. Or, qualitativement FGADR correspond à la base avec le plus fort contraste comparativement aux autres (observable sur les images des figures 4.2 et C.2). Le détail des résultats complets (par lésion) est fourni dans le tableau C.1. Ces résultats très contrastés ne permettent pas de conclure de manière forte sur l'utilité ou non du prétraitement. En revanche, il apparaît deux cas étonnants d'échec complet de segmentation lors du prétraitement appliqué sur la base d'apprentissage MESSIDOR et DDR pour la segmentation des lésions CWS et MA respectivement. Cette instabilité lors de l'apprentissage est inexpliquée.

TABLEAU C.1 Analyse des performances du modèle par lésion avec et sans prétraitement des données. Il en ressort des résultats très contrastés vis-à-vis de la pertinence de ce prétraitement. En particulier, pour deux modèles et sur deux lésions distinctes (résultats indiqués en rouge), le prétraitement a entraîné un échec complet (et inexpliqué) de la segmentation.

Apprentissage	Bases Test	MA		HEM		CWS		EX	
		Sans	Avec	Sans	Avec	Sans	Avec	Sans	Avec
IDRID	IDRID	0.49	0.49	0.52	0.58	0.63	0.67	0.79	0.82
	MESSIDOR	0.47	0.47	0.38	0.39	0.21	0.25	0.50	0.48
	DDR	0.18	0.22	0.20	0.25	0.21	0.23	0.44	0.40
	RET-LES	0.29	0.31	0.23	0.22	0.14	0.06	0.39	0.41
	FGADR	0.05	0.07	0.18	0.31	0.14	0.19	0.19	0.45
MESSIDOR	IDRID	0.42	0.43	0.51	0.41	0.37	0.00	0.60	0.61
	MESSIDOR	0.56	0.56	0.56	0.57	0.14	0.00	0.65	0.58
	DDR	0.31	0.34	0.33	0.33	0.18	0.00	0.37	0.36
	RET-LES	0.22	0.23	0.36	0.35	0.22	0.00	0.40	0.25
	FGADR	0.03	0.07	0.27	0.37	0.10	0.00	0.14	0.30
DDR	IDRID	0.39	0.00	0.54	0.52	0.64	0.45	0.74	0.71
	MESSIDOR	0.48	0.00	0.35	0.51	0.33	0.33	0.55	0.56
	DDR	0.29	0.00	0.52	0.48	0.46	0.41	0.60	0.54
	RET-LES	0.29	0.00	0.41	0.37	0.31	0.27	0.49	0.47
	FGADR	0.08	0.00	0.41	0.45	0.35	0.26	0.38	0.50
RET-LES	IDRID	0.20	0.18	0.53	0.46	0.35	0.20	0.54	0.52
	MESSIDOR	0.22	0.21	0.45	0.45	0.15	0.18	0.35	0.29
	DDR	0.24	0.22	0.32	0.33	0.30	0.32	0.44	0.44
	RET-LES	0.43	0.43	0.56	0.51	0.49	0.44	0.70	0.66
	FGADR	0.12	0.09	0.35	0.36	0.21	0.28	0.34	0.39
FGADR	IDRID	0.36	0.29	0.48	0.57	0.40	0.64	0.73	0.77
	MESSIDOR	0.25	0.46	0.48	0.58	0.24	0.32	0.22	0.54
	DDR	0.21	0.17	0.40	0.42	0.25	0.30	0.21	0.48
	RET-LES	0.36	0.37	0.43	0.37	0.25	0.29	0.31	0.47
	FGADR	0.22	0.23	0.65	0.61	0.42	0.40	0.65	0.64
ALL	IDRID	0.47	0.45	0.67	0.60	0.73	0.61	0.81	0.79
	MESSIDOR	0.51	0.51	0.52	0.58	0.33	0.41	0.59	0.58
	DDR	0.24	0.28	0.45	0.47	0.47	0.42	0.59	0.55
	RET-LES	0.43	0.43	0.57	0.51	0.53	0.45	0.70	0.68
	FGADR	0.24	0.21	0.66	0.61	0.45	0.39	0.65	0.64

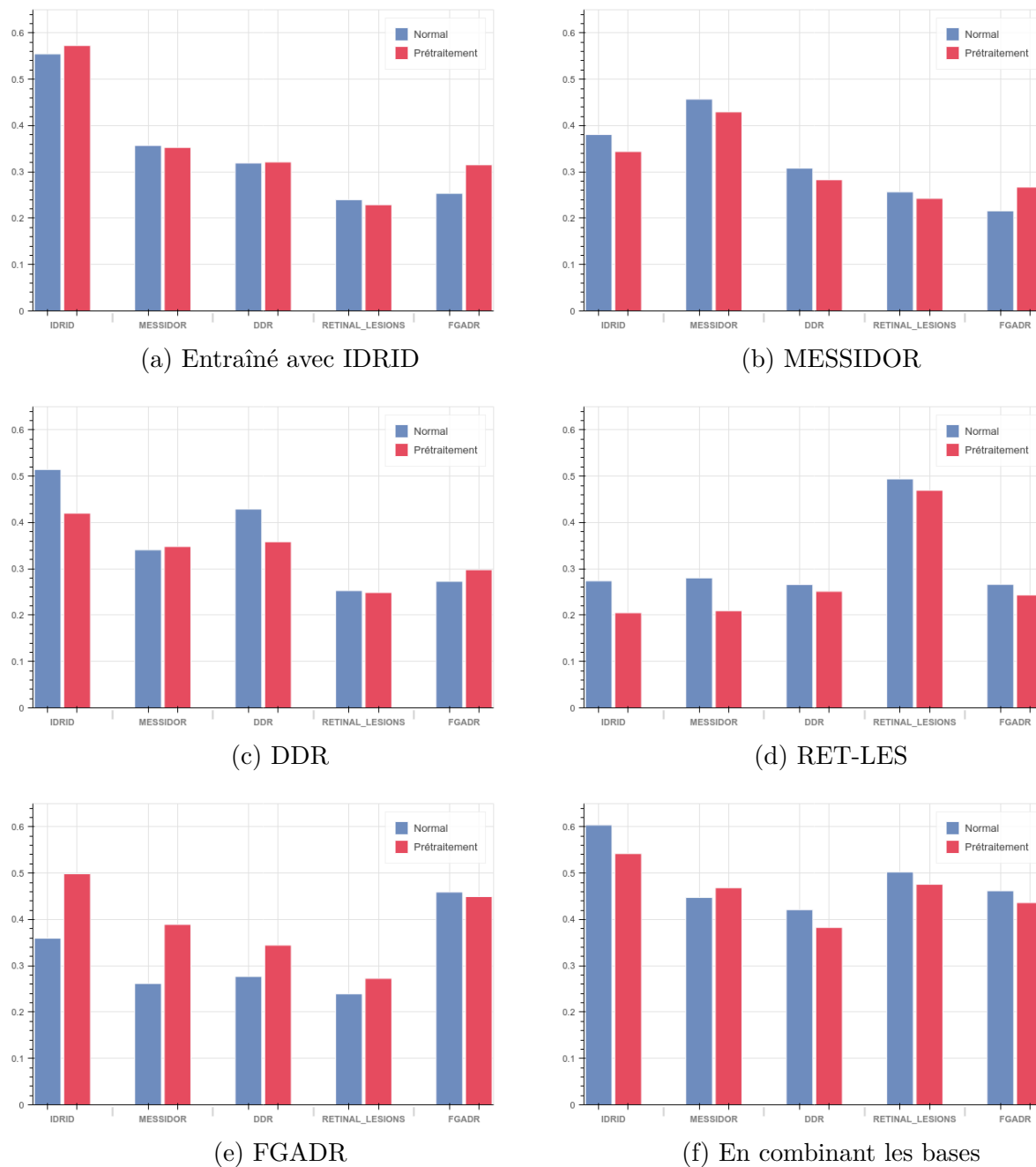


FIGURE C.4 Performance (mIoU) comparée avec et sans pré-traitement sur les différents ensemble de test, en fonction de la base d'entraînement

C.2 La détection, mesure alternative à la segmentation

Nous avons analysé les modèles à travers leur performance en terme de segmentation sémantique. Les métriques employées pour cela se basent toute sur une analyse pixel-par-pixel : un

pixel est soit correctement classifié, soit il ne l'est pas. Or, il existe en réalité une ambiguïté sur la nature de chacun des pixels d'une annotation, ambiguïté qui provient soit d'une interprétation sur la classe d'une structure qui diffère entre les experts-annotateurs, soit d'une appréciation forcément subjective et incertaine des frontières de la structure annotée. Le réseau, entraîné sur une base de donnée, tend à mimer la subjectivité du style d'annotations, notamment en ce qui concerne l'appréciation de la frontière des structures (tel que montré dans la section 4.3.2). Cela explique en grande partie les performances de segmentation très variables en fonction de la base sur laquelle le réseau est testé. Or, la forme exacte d'une lésion ne fait pas partie des critères de diagnostic de la maladie sous-jacente (du moins pour la rétinopathie diabétique) ; l'optimisation de métriques axées sur la segmentation est donc discutable d'un point de vue clinique. Nous proposons en complément une mesure basée sur la détection des lésions en tant que composantes connectées. Pour chaque classe, s'il existe une intersection non vide entre une composante de la prédiction et de l'annotation, alors celle-ci est comptée comme vraie positive. S'il n'y a pas d'intersection, toutes les composantes concernées de l'annotation sont comptées comme fausses négatives ; inversement, toutes les composantes de la prédiction sont elles comptées comme fausse positives. Ces simples règles sont résumées sur le schéma de la figure C.5.

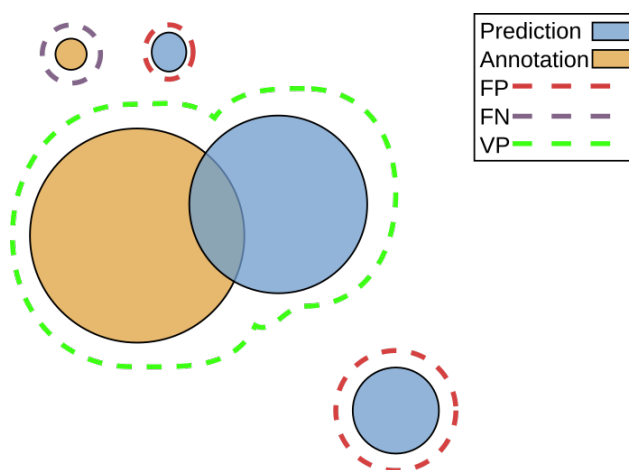


FIGURE C.5 Règles utilisées pour nos métriques de détection. Les contours pointillés indiquent comment sont interprétées les structures suivant l'intersection de la prédiction et de l'annotation : FP : Faux Positifs (2), FN : Faux Négatifs (1), VP : Vrais Positifs (1)

En construisant ainsi une matrice de confusion multiclasses, il est possible de dériver l'équivalent des mesures de sensibilité et de précision, ainsi qu'un score F1, c'est-à-dire la moyenne harmonique des deux. Notons qu'en revanche, en l'absence d'une notion de « Vrai Négatif », il n'est pas possible de mesurer la spécificité.

Ces métriques ne peuvent pas remplacer à elles seules l’analyse par segmentation. En effet, dans le cas d’un modèle qui prédirait qu’une prédiction uniforme sur toute l’image (et donc une seule composante connectée), serait considéré comme idéal pour la classe considérée. Il s’agit donc simplement d’un outils d’analyse complémentaire à la segmentation pour comprendre la mécanique des modèles entraînés.

Les résultats sont donnés sur les graphiques de la figure C.6. On y constate qu’à nouveau, FGADR apparaît être la base de données qui pose le plus de problème aux divers modèles. Ainsi, le modèle entraîné sur cette base voit sa sensibilité chuter drastiquement lorsqu’il est testé sur des données issues d’autres bases (indiquant donc un grand nombre de lésions non détectées). L’effet inverse s’observe également lors qu’un modèle est entraîné sur IDRiD ou MESSIDOR (par ailleurs deux bases relativement petites), la précision s’écroule lorsqu’il est testé sur FGADR. Il est intéressant de noter également que le modèle entraîné sur RETLES, alors qu’il possédait le plus faible mIoU global en terme de segmentation (tableau 4.4) se révèle être le meilleur modèle en terme de précision et score F1 pour la détection. Cela n’a cependant rien d’étonnant quand on considère que ce modèle segmente peu d’instances et regroupe les lésions en de larges structures peu précises. En revanche, le modèle $\mathcal{M}_S$ conserve sa place en haut du palmarès en terme de sensibilité (et second en terme de F1 et de précision) : en cela, les résultats de détection sont en adéquation avec ceux de segmentation, ce qui conforte l’idée que ce modèle est bien le plus performant et cela de façon relativement indépendante de la métrique considérée.

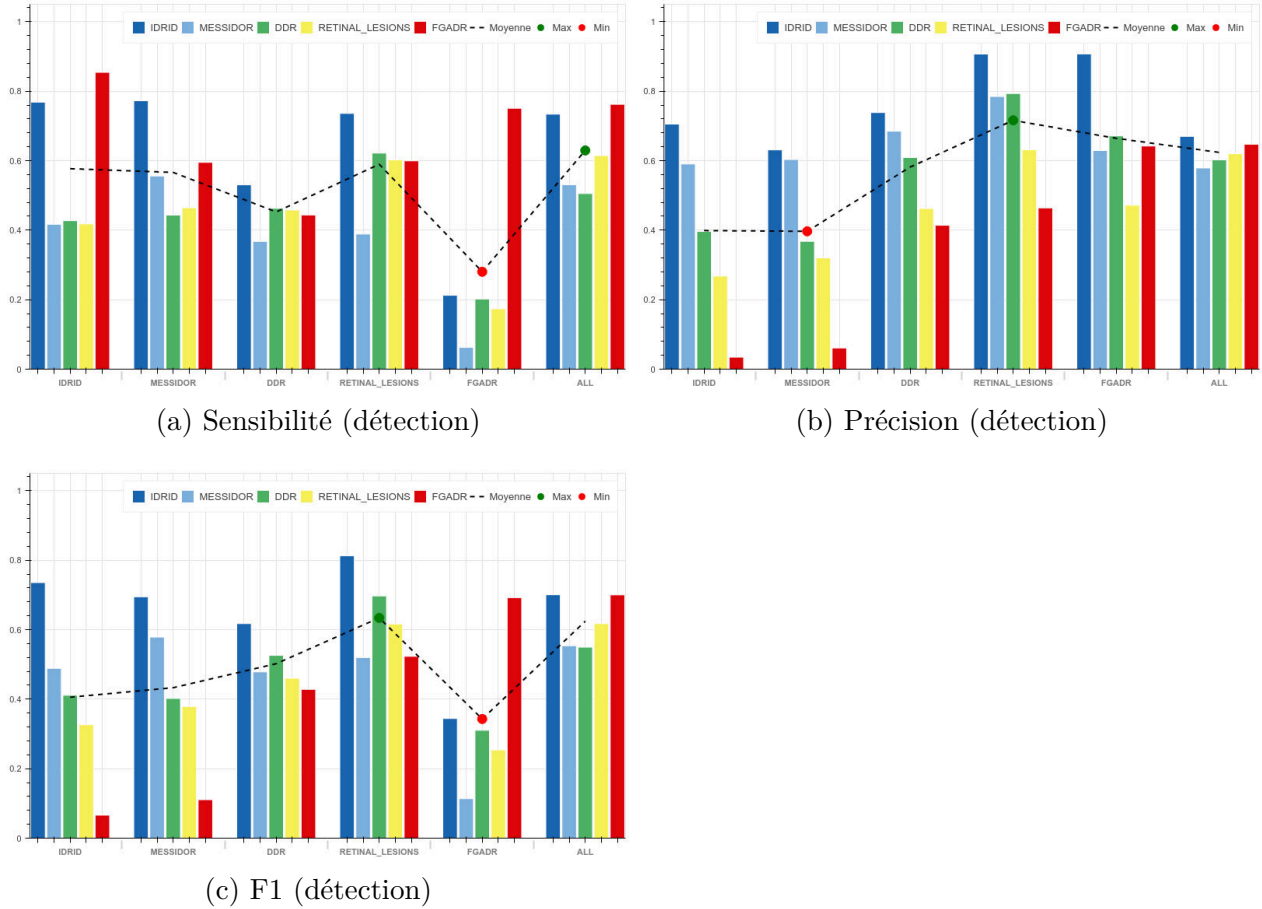


FIGURE C.6 Métriques de détection. Les abscisses représentent les données utilisées pour l'apprentissage et les différentes barres de couleur le score obtenu sur l'ensemble d'évaluation correspondant.

C.3 Conception d'un modèle de segmentation généraliste

Au cours de l'étude sur la généralisation multi-annotations, nous avons expérimenté plusieurs approches pour améliorer les performances du modèle généraliste. Rappelons que la référence que nous utilisons dans nos expériences est $\mathcal{M}_{\mathcal{S}} = \mathcal{M}[\bigcup_i^T \mathcal{B}^{(i)}]$, c'est-à-dire le modèle simplement entraîné sur la combinaison des bases ($T = \{I, M, D, R, F\}$). En premier lieu, la combinaison de tous les ensembles peut sembler contre-intuitive, dans la mesure où certains sont plus ou moins « compatibles » entre eux (c'est-à-dire possèdent ou non les mêmes styles d'annotation). Nous nous sommes interrogé sur les sous-ensembles (ou parties) de T qui pourraient être les plus pertinents. Comme $\mathcal{P}(T) \setminus \emptyset$ (l'ensemble des parties non vides de T) est relativement petit ($\text{Card}(\mathcal{P}(T) \setminus \emptyset) = 2^5 - 1 = 31$), il est relativement aisé d'expérimenter avec l'intégralité des combinaisons possibles de base de données. Les résultats sont exposés

sur le graphique de la figure C.7. On y constate que le meilleur score est bien obtenu par $\mathcal{M}_S$ (tout à droite sur le graphique). La quantité de données prévaut donc sur la qualité de l'homogénéité de celles-ci.

Au lieu de procéder par la concaténation des images au sein d'une même grande base d'apprentissage, il existe d'autres façons de combiner les données. Nous en avons expérimenté deux :

- L'entraînement d'un modèle par base de données suivi de la moyenne des prédictions réalisées par chaque modèle à l'inférence. Cette stratégie est appelée « Ensemble de Modèles ».
- Similairement, l'entraînement d'un modèle par base de données, mais cette fois, la moyenne des poids de chaque modèle est réalisé à l'inférence. Cette idée est empruntée à Wortsman et al. [255] qui la nomment « Soupe de Modèles ».

Notons que dans un cas comme dans l'autre, il suffit de n'entraîner que cinq modèles ($\mathcal{M}[\mathcal{B}^{(I)}]$, $\mathcal{M}[\mathcal{B}^{(M)}]$, $\mathcal{M}[\mathcal{B}^{(D)}]$, $\mathcal{M}[\mathcal{B}^{(R)}]$ et $\mathcal{M}[\mathcal{B}^{(F)}]$), qui seront ensuite combinés aux besoins. À nouveau, 31 combinaisons sont possibles. Précisons cependant que la technique de Wortsman et al. est très récente et n'a été appliquée que sur des réseaux de classification. Pour la segmentation, nous n'avons obtenu de résultats fonctionnels qu'en limitant les « soupes » aux poids de l'encodeur. Pour chacune de celles-ci, 5 décodeurs peuvent donc être adoptés, créant donc $5 \times 31 = 155$ modèles possibles au total. En pratique, dans nos expériences, nous nous sommes limités qu'à 31 modèles (soupe d'encodeurs et décodeur issu de $\mathcal{M}_S$). Les résultats obtenus avec les différentes techniques sont donnés sur la figure C.8. Il y apparaît clairement que la concaténation des données d'entraînement est la plus efficace des trois stratégies, quelle que soit la combinaison empruntée. Nous avons mené d'autres expériences dans le but de combiner différemment les modèles, mais nos efforts en ce sens se sont révélés infructueux (absence de convergence ou performances très en deçà de celles obtenues avec les modèles classiques). En vrac, citons :

- L'adaptation adversarial par un système à double couple générateur/discriminateur (aussi appelé GAN) : le premier GAN est chargé de modifier l'apparence de chaque image d'entrée vers un domaine cible (en l'occurrence $\mathcal{B}^{(I)}$). Le second GAN a pour objectif de modifier le style d'annotation vers celui du domaine cible, là encore $\mathcal{B}^{(I)}$. Entre ces deux GAN, un réseau de segmentation classique manipule les images et les annotations adversarialement converties par chaque GAN.
- La régularisation par apprentissage auto-entraîné (*self training*), en exploitant une large base de données non-labellisées. Un premier réseau est entraîné sur $\mathcal{B}^{(I)}$ et infère sur une grande base d'image (EyePACS). On utilise alors ces prédictions pour entraîner un second réseau, conjointement avec $\mathcal{B}^{(I)}$. Nous n'avons pas obtenu d'améliorations

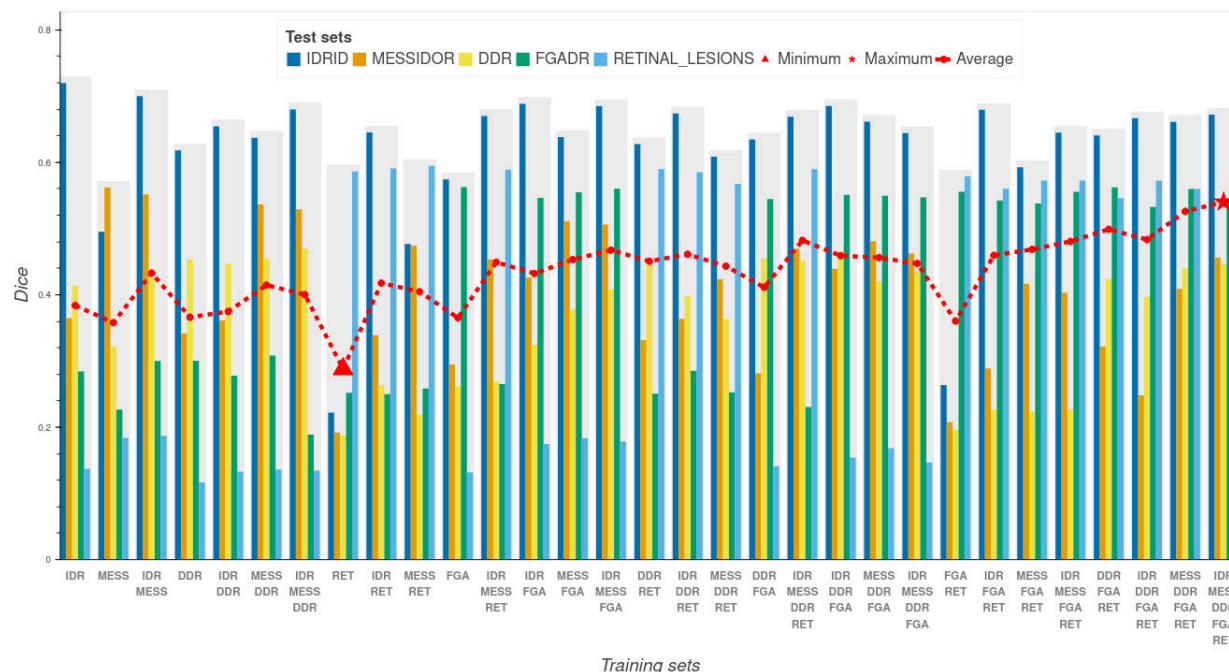


FIGURE C.7 Performance des différents modèles, entraînés chacun sur une partie de T . Les parties sont triées par ordre croissant de cardinalité, c'est-à-dire par le nombre d'images qu'elles contiennent.

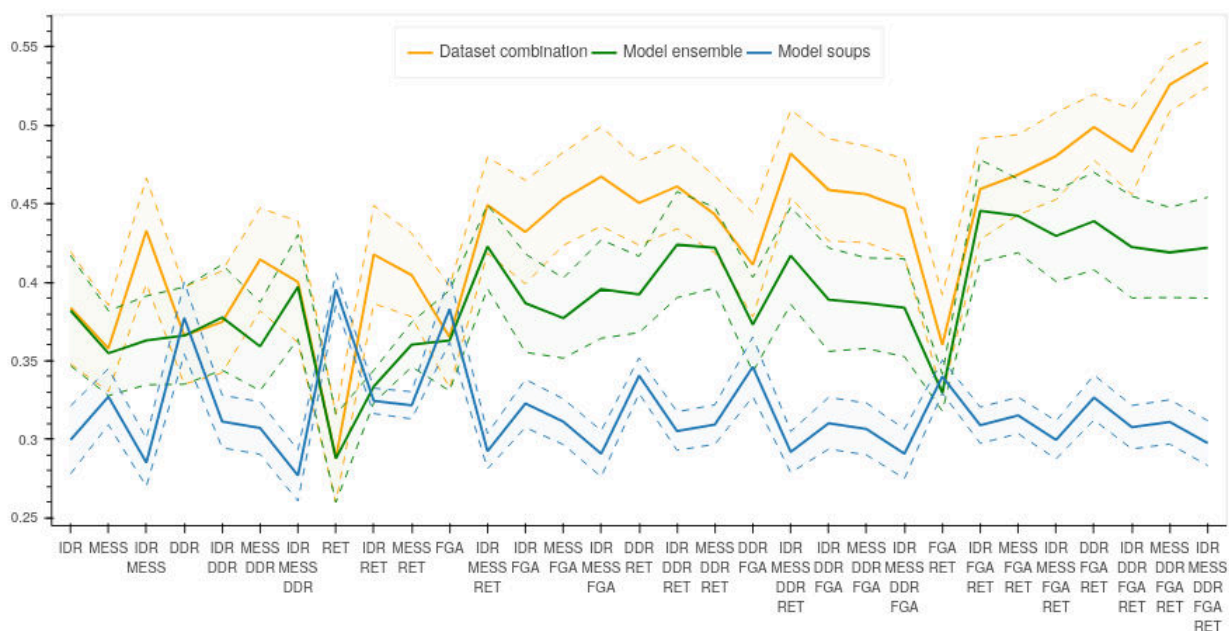


FIGURE C.8 Résultats de différentes méthodes de combinaisons des bases de données. Seul le Dice moyen est fourni ($\frac{1}{5} \sum_{i=1}^5 D(\mathcal{M}, \mathcal{B}_*^{(i)})$). L'écart-type est représenté en pointillés ($\pm \sigma$).

par l'ajout des images auto-annotées.

- La régularisation par apprentissage contrastif (*Contrastive Learning*) en ré-implémentant l'approche de Zhao et al. [256]. Nous n'avons pas réussi à faire converger ce modèle.

Même sans rentrer dans les détails, ces différentes expérimentations illustrent le potentiel de sophistication que peut prendre la segmentation sémantique du fond d'œil. Chacune de ces expériences semble théoriquement fondée ; malheureusement en pratique, nous nous sommes heurté à la difficulté de les faire converger adéquatement, y compris en tentant de respecter scrupuleusement les méthodologies existantes. Si cela nous confronte à la difficulté de la reproductibilité de la littérature, il n'est pas exclu que des efforts plus conséquents dans une voie donnée pourraient mener à des résultats bien plus aboutis.