Intelligence artificielle pour la caractérisation du cancer de la prostate par agressivité en IRM multiparamétrique

Audrey Duran

N°d'ordre NNT : 2022LYSEI008

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
(Institut National des Sciences Appliquées, INSA - Lyon)

Ecole Doctorale N° 160
(ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE)

Spécialité/ discipline de doctorat :
Traitement du Signal et de l'Image

Soutenue publiquement le 03/02/2022, par :
Audrey Duran

Intelligence artificielle pour la caractérisation du cancer de la prostate par agressivité en IRM multiparamétrique

---

Devant le jury composé de :

Mériaudeau, Fabrice	Professeur des Universités, Université de Bourgogne	Rapporteur
Petitjean, Caroline	Professeure des Universités, Université de Rouen	Rapporteure
Jodoin, Pierre-Marc	Professeur des Universités, Université de Sherbrooke	Examinateur
Renard-Penna, Raphaële	Professeure-Praticien Hospitalier, Sorbonne Universités	Examinatrice
Rouvière, Olivier	Professeur-Praticien Hospitalier, Université Lyon 1	Examinateur
Lartizien, Carole	Directrice de recherche, CNRS, Lyon	Directrice de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **https://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **https://edeea.universite-lyon.fr** Sec. : Stéphanie CAUVIN Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 **philippe.delachartre@insa-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**M. Philippe NORMAND** Université Claude Bernard Lyon 1 UMR 5557 Lab. d'Ecologie Microbienne Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://ediss.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 **stephane.benayoun@ec-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **https://edsciencessociales.universite-lyon.fr** Sec. : Mélina FAVETON INSA : J.Y. TOUSSAINT Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Christian MONTES** Université Lumière Lyon 2 86 Rue Pasteur 69365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Résumé

Le cancer de la prostate (CaP) est le cancer le plus diagnostiqué dans plus de la moitié des pays du monde et le cinquième cancer le plus meurtrier chez les hommes en 2020. Le diagnostic du CaP inclut l'acquisition d'une imagerie par résonance magnétique multiparamétrique (IRM-mp) - qui combine une séquence T2-pondérée (T2-w), une imagerie pondérée en diffusion (DWI) et une séquence dynamique de contraste amélioré (DCE) - avant même la réalisation de biopsies. L'analyse jointe de ces images multimodales est fastidieuse et chronophage, en particulier lorsque les séquences mènent à des conclusions différentes. En outre, la sensibilité de l'IRM reste faible pour les cancers peu agressifs et la variabilité inter-observateur élevée. De plus, l'analyse visuelle ne permet pas aujourd'hui de déterminer l'agressivité des cancers, caractérisée par le score de Gleason (GS).

C'est pourquoi de nombreux systèmes d'aide à la détection et au diagnostic (CAD) basés sur des modèles statistiques par apprentissage ont été proposés ces dernières années, dans le but d'assister les radiologues dans leur diagnostic. Toutefois, la majorité de ces systèmes se concentrent sur une tâche de détection binaire des lésions cliniquement significatives (CS). L'objectif de cette thèse est d'élaborer un système CAD pour détecter les CaP sur des images IRM-mp, mais aussi de caractériser leur agressivité en prédisant le GS associé.

Dans une première partie, nous présentons un système CAD supervisé permettant de segmenter le CaP par agressivité à partir des cartes T2-w et ADC. Ce réseau de neurones multiclasse entraîné de bout en bout segmente simultanément la prostate et les lésions par agressivité. Après avoir encodé l'information dans un espace latent, le réseau est séparé en deux branches : 1) la première segmente la prostate 2) la seconde utilise cette information anatomique pour la détection et classification par GS des lésions prostatiques. Le modèle a été entraîné et évalué en validation croisée à 5 plis sur une base de données hétérogène de 219 examens IRM acquis sur trois scanners différents avant prostatectomie. Pour la tâche de classification par GS, le coefficient kappa de Cohen quadratiquement pondéré (κ) est de 0.418 ± 0.138, ce qui représente le meilleur kappa par lésions pour une tâche de segmentation par GS à notre connaissance. Le modèle présente également des capacités de généralisation encourageantes sur le jeu de données public PROSTATEx-2. Dans une deuxième partie, nous nous penchons sur un modèle faiblement supervisé, permettant l'inclusion de données où les lésions sont identifiées par des points seulement, pour un gain de temps conséquent et l'inclusion de bases de données établies sur la biopsie. Concernant la tâche de classification par GS, nous montrons que les performances approchent celles obtenues avec le modèle totalement supervisé de référence, en n'ayant que 6% de voxels annotés pour l'entraînement. Dans une dernière partie, nous étudions l'apport de l'imagerie DCE, séquence souvent omise en entrée des modèles profonds, pour la détection et la caractérisation du CaP. Plusieurs stratégies d'encodage de l'information dynamique dans une architecture de type U-Net sont étudiées. Nous montrons que les cartes paramétriques dérivées des examens IRM DCE ont un impact positif sur les performances de segmentation et de classification du CaP.

Abstract

Prostate cancer (PCa) is the most frequently diagnosed cancer in men in more than half the countries in the world and the fifth leading cause of cancer death among men in 2020. Diagnosis of PCa includes multiparametric magnetic resonance imaging acquisition (mp-MRI) - which combines T2 weighted (T2-w), diffusion weighted imaging (DWI) and dynamic contrast enhanced (DCE) sequences - prior to any biopsy. The joint analysis of these multimodal images is time demanding and challenging, especially when individual MR sequences yield conflicting findings. In addition, the sensitivity of MRI is low for less aggressive cancers and inter-reader reproducibility remains moderate at best. Moreover, visual analysis does not currently allow to determine the cancer aggressiveness, characterized by the Gleason score (GS).

This is why computer-aided detection and diagnosis (CAD) systems based on statistical learning models have been proposed in recent years, to assist radiologists in their diagnostic task, but the vast majority of these models focus on the binary detection of clinically significant (CS) lesions. The objective of this thesis is to develop a CAD system to detect and segment PCa on mp-MRI images but also to characterize their aggressiveness, by predicting the associated GS.

In a first part, we present a supervised CAD system to segment PCa by aggressiveness from T2-w and ADC maps. This end-to-end multi-class neural network jointly segments the prostate gland and cancer lesions with GS group grading. After encoding the information on a latent space, the network is separated in two branches : 1) the first branch performs prostate segmentation 2) the second branch uses this zonal prior as an attention gate for the detection and grading of prostate lesions. The model was trained and validated with a 5-fold cross-validation on a heterogeneous series of 219 MRI exams acquired on three different scanners prior prostatectomy. Regarding the automatic GS group grading, Cohen's quadratic weighted kappa coefficient (κ) is 0.418 ± 0.138, which is the best reported lesion-wise kappa for GS segmentation to our knowledge. The model has also encouraging generalization capacities on the PROSTATEx-2 public dataset. In a second part, we focus on a weakly supervised model that allows the inclusion of partly annotated data, where the lesions are identified by points only, for a consequent saving of time and the inclusion of biopsy-based databases. Regarding the automatic GS group grading on our private dataset, we show that we can approach performance achieved with the baseline fully supervised model while considering 6% of annotated voxels only for training. In the last part, we study the contribution of DCE MRI, a sequence often omitted as input to deep models, for the detection and characterization of PCa. We evaluate several ways to encode the perfusion from the DCE MRI information in a U-Net like architecture. Parametric maps derived from DCE MR exams are shown to positively impact segmentation and grading performance of PCa lesions.

Remerciements

Je tiens tout d'abord à remercier les rapporteur·e·s Caroline Petitjean et Fabrice Mériaudeau pour leur lecture attentive de mon manuscrit. Merci à Pierre-Marc Jodoin, Raphaële Renard-Penna et Olivier Rouvière d'avoir pris de leur temps précieux pour évaluer ma thèse et partager leur expertise.

Merci en particulier à Pierre-Marc Jodoin, pour les 3 mois passés au sein de son laboratoire VITAL à Sherbrooke. Ses enseignements ont été précieux pour la suite. Merci aux membres du laboratoire VITAL pour leur accueil et leurs revues de code.

Merci également à Olivier Rouvière pour toutes les séances de lecture et les réponses rapides aux nombreuses questions posées.

Merci à Carole Lartizien, ma directrice de thèse, de m'avoir permis de réaliser cette thèse qui sortait de ma spécialité initiale. Merci pour l'encadrement sérieux et régulier pendant ces 3 années.

Merci à Michaël Sdika et Odyssée Merveille, permanent·e·s à CREATIS, pour leurs précieux conseils avant la soutenance.

Merci aux collègues doctorant·e·s, pour les échanges enrichissants que ce soit en pause, en soirée, à Domus ou autour de mots fléchés : Nono, Maxime, Arthur, Habib, Pilar, Benoît four, Charly, Ali, Nina, Florian, Suzanne, Antonio, Ludmilla, Louise, Théo, Pierre-Jean, Sébastien, Sophie, Frank, Cyril, Anne-Lise, Benoît, Charles, Raoul, Pierre.

Merci à Gaspard, meilleur (et seul) stagiaire ever.

Merci au groupe L.I.F.E. initié au laboratoire, qui oeuvre pour le climat.

Merci à tous les membres du laboratoire CREATIS avec qui j'ai pu échanger durant ces 3 années, qui ont contribué à rendre cette expérience aussi riche.

Je tiens à remercier deepl pour sa réactivité et pertinence malgré quelques égarements.

Merci à mes amis et à ma famille d'être là. Merci aux BIMs, au CLAV, au Groupe 81, au comeback et tous les autres.

Un immense merci à Sam, qui m'a accompagnée ces trois années de thèse et peut maintenant parler de score de Gleason.

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

ADC Coefficient apparent de diffusion (*Apparent Diffusion Coefficient*)

AUC Aire sous la courbe (*Area Under Curve)*

CAD Système d'aide au diagnostic (*Computer-Aided Diagnosis*)

CADe Système d'aide à la détection

CADx Système d'aide à la décision

CAM Cartes d'activation de classe (*Class Activation Maps*)

CaP Cancer de la Prostate

CLARA-P CorréLations Anatomo-RAdiologiques en IRM de Prostate, base de données utilisée dans cette thèse

CNN Réseau de neurones convolutionnel (*Convolutional Neural Network*)

CREATIS Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé

CRF Champ aléatoire conditionnel (*Conditional Random Field*)

CS Cliniquement Significatif

D Dimension

DCE Imagerie de perfusion (*Dynamic Contrast Enhanced*)

DWI Imagerie pondérée en diffusion (*Diffusion Weighted Imaging*)

FN Faux Négatif

FOV Champ de vue (*Field Of View*)

FP Faux Positif

FROC *Free-response Receiver Operating Characteristics*

GAN Réseaux adversaires génératifs (*Generative adversarial network*)

GAP *Global Average Pooling*

GG Groupe de Gleason

GS Score de Gleason (*Gleason score*)

HIFU Ultrasons focalisés de haute intensité (*High-Intensity Focused Ultrasound*)

IoU Intersection sur l'union

IRM Imagerie par Résonance Magnétique

IRM-mp IRM multi-paramétrique

IRM-bp IRM bi-paramétrique

ISUP Société internationale de pathologie urologique (*International Society of Urological Pathology*). Cet acronyme est utilisé pour faire référence au classement par groupe de Gleason proposé lors d'une conférence de consensus.

PERFUSE ***Per**sonalized **F**ocused **U**ltrasound **S**urg**e**r**y** of Localized Prostate Cancer*, projet de l'Agence nationale de la recherche (ANR) financeur de cette thèse.

PI-RADS *Prostate Imaging–Reporting and Data System*

PSA Antigènes spécifiques à la prostate (*Prostatic Specific Antigen)*

PT Prostatectomie Totale

RHU Recherche Hospitalo-Universitaire

ROC *Receiver Operating Characteristics*

RT Radiothérapie

RTE Radiothérapie Externe

T Tesla

$\mathbf{T}_E$ Temps d'écho

$\mathbf{T}_R$ Temps de répétition

TTP Temps au pic (*Time To Peak*)

T1-w Imagerie T1-pondérée (*T1-weighted*)

T2-w Imagerie T2-pondérée (*T2-weighted*)

VN Vrai Négatif

VP Vrai Positif

ZC Zone Centrale

ZP Zone Périphérique

ZT Zone de Transition

Chapitre 1

Introduction

Introduction générale

Le cancer de la prostate (CaP) est le cancer le plus diagnostiqué chez l'homme dans plus de la moitié des pays du monde [118]. Avec plus de 1,4 million de nouveaux cas et 375 000 décès dans le monde, c'est le second cancer le plus fréquent et le cinquième cancer le plus meurtrier chez les hommes en 2020.

L'imagerie par résonance magnétique (IRM) fait dorénavant partie du processus de diagnostic classique du CaP. Elle est multiparamétrique (IRM-mp) et combine une séquence anatomique T2-w à trois dimensions (noté 3D), une séquence de diffusion DWI 3D également et une séquence dynamique ou de perfusion DCE à 4D. En conséquence, le volume de données à analyser conjointement est considérable, rendant la tâche complexe, d'autant plus lorsque les séquences mènent à des conclusions différentes. Il en résulte une variabilité inter-observateur élevée, les radiologues peu expérimentés obtenant des performances en deçà de celles des médecins entraînés à cet exercice.

Pour toutes ces raisons, la recherche autour des systèmes d'aide au diagnostic (CAD) pour assister les radiologues à analyser les IRM-mp est prolifique depuis plus d'une dizaine d'années. On peut distinguer les approches radiomiques, où des paramètres quantitatifs sont extraits des images radiologiques, des approches basées sur l'apprentissage, en particulier l'apprentissage profond. Dans ce cas, le modèle extrait les caractéristiques discriminantes sans sélection préalable. Dans cette thèse, nous avons choisi d'étudier les méthodes par apprentissage profond pour l'analyse des données d'imagerie IRM de la prostate, plus particulièrement pour la détection et la segmentation par agressivité des cancers des zones périphérique et de transition.

Les verrous à lever en CAD pour le CaP

Les systèmes CAD pour la caractérisation du cancer de la prostate sont étudiés depuis plus d'une dizaine d'années (voir chapitre 4). Toutefois, le problème est loin d'être résolu et les points suivants méritent d'être approfondis :

— Tout d'abord, les performances restent à améliorer. La plupart des systèmes restent moins performants que les radiologues et présentent donc encore une marge de progression.

— En outre, les modèles sont souvent validés sur des bases relativement petites et souvent homogènes, sans test sur une base externe. Ce problème de généralisation du modèle est un point critique pour une utilisation ultérieure en clinique, d'autant plus qu'une utilisation dans un autre centre, où des machines différentes sont utilisées pour l'acquisition, induit un changement de domaine

auquel les réseaux de neurones profonds sont souvent sensibles. Néanmoins, le manque de bases de données publiques suffisamment étoffées rend cette validation difficile.

— Ensuite, un des points clés pour l'établissement d'un système robuste concerne la base de données utilisée pour l'entraînement et la validation du modèle. Or, l'annotation des données par un expert est une tâche extrêmement chronophage et fastidieuse, qui peut prendre plusieurs années pour obtenir une base de taille correcte, ce qui limite l'élaboration de bases de données conséquentes. L'utilisation de données sans ou avec peu d'annotations peut permettre de remédier à ce problème et constitue un axe de recherche de fort intérêt.

— Enfin, au commencement de cette thèse, aucune étude n'avait été faite sur la caractérisation de l'agressivité du cancer détecté sur les IRM. Or, l'agressivité du cancer est une information cruciale dont va dépendre le choix thérapeutique. La détermination du score de Gleason d'une lésion est actuellement réalisée par l'intermédiaire de biopsie, qui, en plus d'être une technique invasive, manque de spécificité et peut sur ou sous-estimer le grade [44]. De plus, dans un contexte où la surveillance active est préconisée lorsqu'un cancer de faible agressivité a été identifié, pouvoir suivre l'évolution du cancer par une méthode non invasive telle que l'IRM est un enjeu crucial.

Contexte et motivations

Le travail présenté dans cette thèse a été réalisé au laboratoire CREATIS (Centre de Recherche En Acquisition et Traitement de l'Image pour la Santé, www.creatis.insa-lyon.fr), à Lyon. Ce travail s'inscrit dans un projet de Recherche Hospitalo-Universitaire (RHU) en santé nommé PERFUSE pour ***Per**sonalized **F**ocused **U**ltrasound **S**urgery* (www.rhu-perfuse.fr). Le but de ce projet de 6 ans (2017-2023) est de mettre en synergie l'expertise académique de chercheurs (LabTAU et CREATIS), urologues, oncologues, radiologues et biologistes (Hospices Civils de Lyon et Centre Léon Bérard), avec les compétences de deux compagnies privées EDAP et VERMON, spécialisées dans les technologies médicales par ultrason. Il a pour objectif d'évaluer les résultats oncologiques du traitement focal HIFU (décrit section 2.6) du cancer de la prostate et de proposer un ensemble cohérent d'innovations pour permettre un traitement focal personnalisé et plus rationnel, du diagnostic à une technique optimale du traitement HIFU.

La mise en place du traitement HIFU implique une localisation des foyers malins dans la glande prostatique à partir de l'imagerie ou plus précisément de l'IRM-mp, qui s'est récemment imposée comme technique de diagnostic de référence (voir section 2.7). Toutefois, l'analyse des séquences IRM reste difficile et présente une forte variabilité intra et inter-observateur. Le développement de systèmes CAD est donc très étudié pour assister le radiologue dans cette tâche de localisation des lésions.

L'objectif de CREATIS dans le projet PERFUSE est donc de proposer un système permettant d'établir une cartographie du cancer dans l'IRM-mp.

Grâce à ce projet, nous avons eu accès à une base de données privée (CLARA-P, présentée dans le chapitre 5), pour laquelle la vérité terrain basée sur la prostatectomie est disponible. Les systèmes CAD par apprentissage que nous avons développés au cours de cette thèse ont été entraînés et évalués sur cette riche base de données, présentant de nombreuses hétérogénéités (constructeur des scanners utilisés pour les acquisitions multiples, paramètres d'acquisition différents, etc.).

Pour évaluer les capacités de généralisation de nos travaux, nous avons également inclus dans l'évaluation la base de données de challenge ProstateX-2.

Plan de la thèse

La première partie de ce manuscrit présente les contextes médical et scientifique de la thèse. Tout d'abord, nous présentons le cancer de la prostate dans le chapitre 2. Ensuite, dans le chapitre 3, nous introduisons les bases de l'apprentissage profond, plus précisément appliqué à des problématiques de segmentation d'images. Enfin, nous dressons dans le chapitre 4 un état de l'art des systèmes CAD pour le CaP en imagerie IRM proposés dans la littérature.

La seconde partie est consacrée à la présentation des données et des choix effectués dans cette thèse dans le chapitre 5 et le chapitre 6 respectivement.

La troisième partie de la thèse expose les contributions. Tout d'abord, nous présentons chapitre 7 un premier système CAD supervisé, robuste face à des données hétérogènes, qui localise et caractérise l'agressivité des lésions à partir du T2-w et des cartes ADC. Toutefois, ce modèle supervisé nécessite une grande quantité de données, difficiles à obtenir et fastidieuses à annoter, d'autant plus lorsque les images doivent être confrontées ultérieurement à la vérité terrain (prostatectomie ou biopsie). Dans le chapitre 8, nous proposons une approche faiblement supervisée, qui permet d'alléger le fardeau de l'annotation des données dans le domaine médical. Enfin, dans le chapitre 9, nous étudions l'apport de l'imagerie dynamique (DCE), rarement incluse dans les systèmes CAD. Nous terminons par les conclusions et perspectives de la thèse dans le chapitre 10.

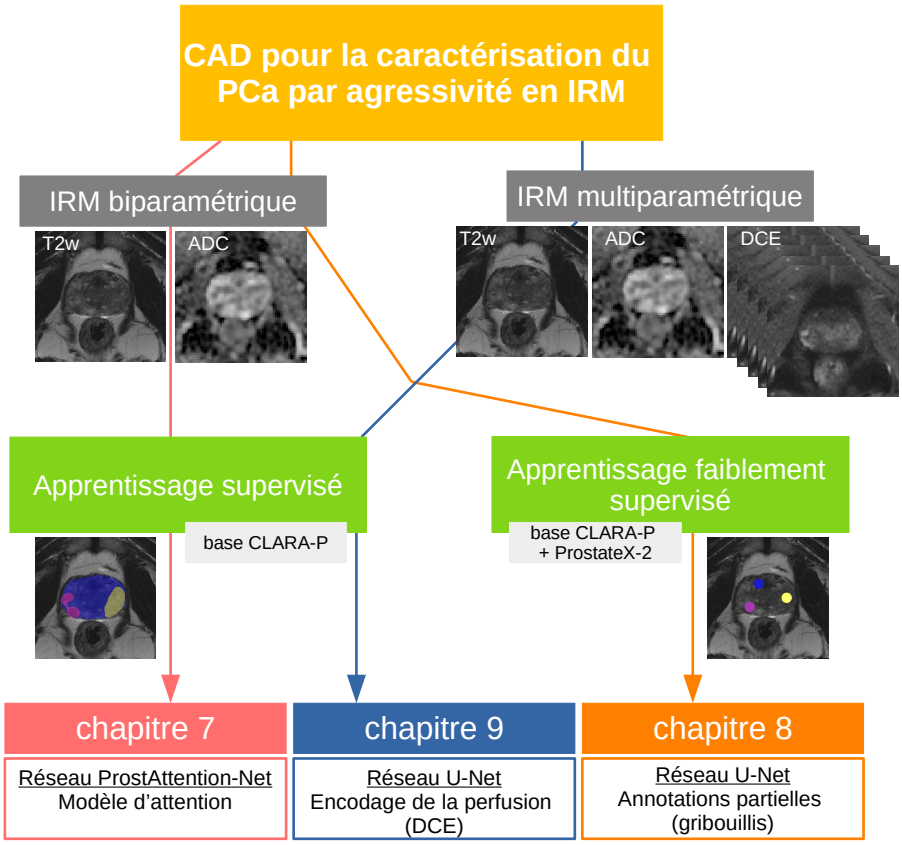


FIGURE 1.1 – Résumé des contributions.

Chapitre 2

Le cancer de la prostate

Sommaire

2.1 Introduction

Dans ce chapitre, nous présentons quelques éléments de l'anatomie et de la fonction de la prostate au sein de l'appareil reproducteur masculin. Nous nous intéressons ensuite au cancer de la prostate en fournissant quelques données épidémiologiques avant de décrire le protocole de diagnostic et prise en charge thérapeutique de cette pathologie. Les parties diagnostic et traitement sont basées sur les dernières recommandations françaises du Comité de cancérologie de l'AFU pour 2020-2022 décrites dans [106].

2.2 Épidémiologie

Le cancer de la prostate (CaP) est le cancer le plus diagnostiqué chez l'homme dans plus de la moitié des pays du monde [118]. Avec plus de 1.4 million de nouveaux cas et 375 000 décès dans le monde, c'est le second cancer le plus fréquent et le cinquième cancer le plus meurtrier chez les hommes en 2020. En France, le cancer de la prostate se situe au 3ème rang des décès par cancer chez l'homme (8 512 décès estimés en 2015 – incidence : 8.9/100 000).

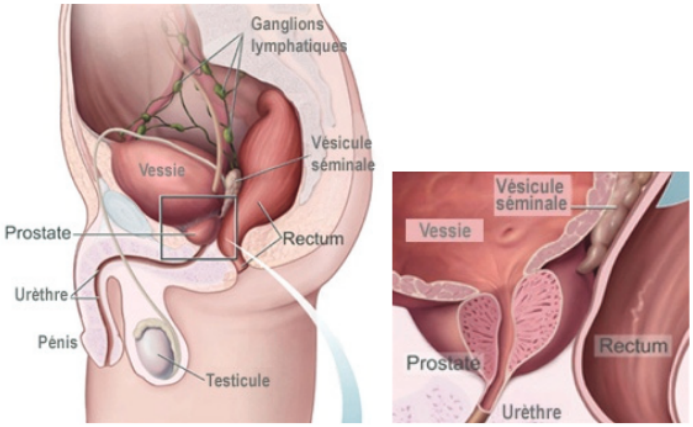


FIGURE 2.1 – Pelvis et appareil reproducteur masculins. Source : Créé par *US government agency National Cancer Institute*, traduction française par F. Lamiot. Domaine public via Wikimedia Commons.

Facteurs de risque Le risque est particulièrement élevé chez les personnes âgées : près de 79 % des décès concernent des hommes de 75 ans et plus avec un âge médian au diagnostic de 68 ans en 2018 [74]. En plus de l'âge, les antécédents familiaux représentent un puissant facteur de risque ; le CaP est le cancer pour lequel le poids de l'hérédité est prépondérant [106]. Enfin, l'ethnie constitue également un facteur de risque, une origine africaine ou afro-caribéenne augmentant les risques de CaP. Aux Antilles françaises - où 90 % de la population est d'origine ethnique d'ascendance africaine - l'incidence du CaP et sa mortalité sont deux fois plus élevées par rapport à la France métropolitaine (respectivement 173/100000 pour l'incidence et 23/100000 pour la mortalité selon le rapport publié par Santé publique France en 2019).

2.3 Anatomie et fonctions de la prostate

La prostate (voir figure 2.1) est une glande de l'appareil génital masculin. Elle est située dans le bassin, sous la vessie en avant du rectum et entoure le début de l'urètre, canal permettant d'éliminer l'urine de la vessie. Une prostate saine a la forme d'une châtaigne d'environ 3 centimètres de hauteur et 4 centimètres de large, elle ne pèse pas plus de 20 grammes à l'âge adulte. La prostate est entourée d'une capsule et formée de plusieurs lobes : un lobe prostatique antérieur, deux lobes latéraux et un lobe médian, aussi appelé lobe de Home. Elle se divise en 3 zones glandulaires principales :

— une zone périphérique (ZP) : c'est la région de la prostate la plus proche du rectum. Elle constitue la plus grande zone de la prostate et en recouvre les faces latérales et postérieure.

— une zone de transition (ZT) : c'est la zone comprenant 2 lobes situés au milieu de la prostate en avant des zones périphérique et centrale. Elle entoure l'urètre et représente environ 5% de la prostate jusqu'à l'âge de 40 ans. Avec le vieillissement, cette zone augmente en taille pour devenir la plus grosse partie de la prostate. C'est ce qu'on appelle un adénome de la prostate (également appelé hypertrophie bénigne de la prostate, noté HBP) qui survient chez presque tous les hommes de plus de 70 ans. L'augmentation de taille de la zone de transition a pour effet de pousser la zone périphérique vers le rectum.

— une zone centrale (ZC) : c'est la partie de la prostate située à la base entourant les canaux éjaculateurs. Cette zone de forme conique est souvent associée à la ZT.

2.4 Caractéristiques du cancer de la prostate

La grande majorité des cancers prostatiques sont situés dans la ZP ; ils représentent environ 75% des cancers, les 25% restant se trouvant dans la ZT.

Score de Gleason

Le score de Gleason (GS) permet de constater le degré d'agressivité du CaP. C'est un facteur pronostique majeur : il permet de mesurer l'étendue et l'agressivité de la maladie et sera considéré pour le choix du traitement à proposer.

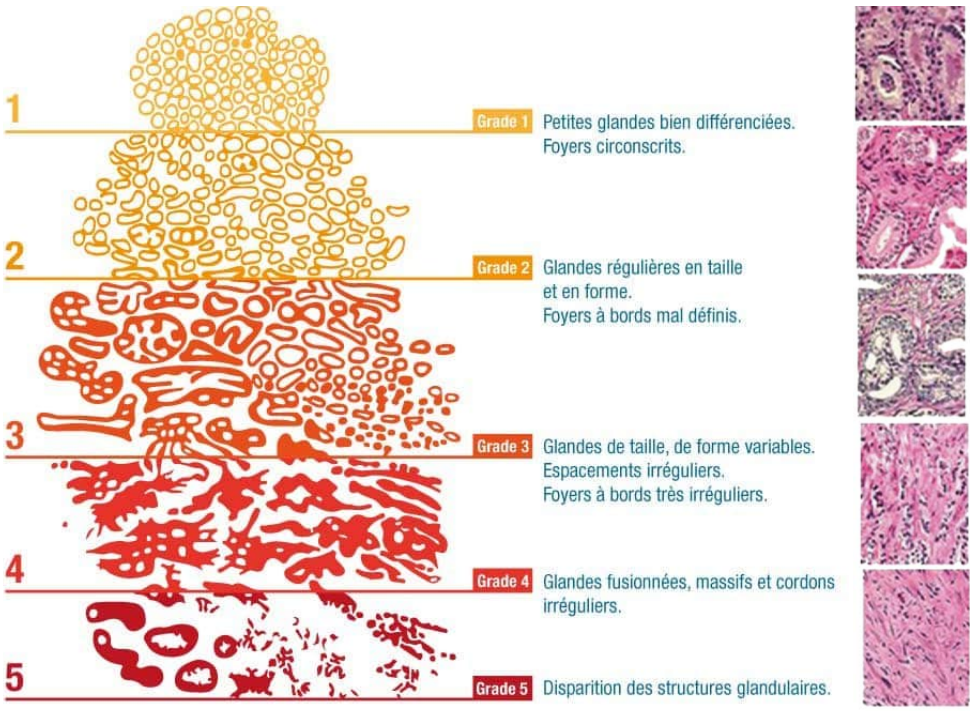


FIGURE 2.2 – Représentation du grade de Gleason en fonction de la différenciation des cellules (schéma et coupe histologique). Source : Prostanet (gauche) et adaptée de l'article Harnden et al., 2007 (droite).

Il est défini à partir de l'analyse de biopsies : en fonction du niveau de différenciation des cellules, on attribue d'abord un grade architectural compris entre 1 et 5 aux cellules composant la tumeur (voir figure 2.2).

Le cancer est souvent hétérogène, des foyers tumoraux d'évolution différente, à des stades de différenciation variables pouvant coexister au sein d'un même tissu prélevé lors de la biopsie. Le score de Gleason est alors obtenu en additionnant les 2 grades histologiques les plus représentés au sein du cancer.

La classification définie par Gleason en 1966 comportait 5 grades architecturaux allant de 1 à 5, dont la somme définissait 9 scores de 2 (1+1) à 10 (5+5). Le grade majoritaire est le premier terme de l'addition : une tumeur de Gleason 7 (3+4) est formé en majorité par le grade 3 et minoritairement par du grade 4, contrairement a

Groupe 1	Score de Gleason 6 (3 + 3)
Groupe 2	Score de Gleason 7 (3 + 4)
Groupe 3	Score de Gleason 7 (4 + 3)
Groupe 4	Score de Gleason 8 (3 + 5, 5 + 3 ou 4 + 4)
Groupe 5	Score de Gleason 9 ou 10

TABLEAU 2.1 – Groupes pronostiques de la classification ISUP 2016

une tumeur de Gleason 7 (4+3) qui sera donc plus agressive.

Ce système a été revu lors des conférences de consensus de l'*International Society of Urological Pathology* (ISUP), en 2005 puis en 2014 [43].

En effet, cette classification comportait plusieurs défauts. Tout d'abord, la presque totalité des CaP diagnostiqués actuellement ont un score minimal de 6, correspondant à des cancers très bien différenciés. Il est de ce fait difficile pour les patients de comprendre qu'ils ont un cancer indolent, alors que leur score se situe dans la médiane de l'échelle de Gleason. De plus, le score de Gleason *stricto sensu* ne fait pas de différences entres les scores 7 (3 + 4) et 7 (4 + 3).

Une nouvelle classification a donc été proposée par l'ISUP [43], avec 5 groupes pronostiques présentés tableau 2.1. Cette classification sert de référence de nos jours.

2.5 Diagnostic du cancer de la prostate

En France, le dépistage systématique du CaP chez un homme asymptomatique n'est pas recommandé par la Haute Autorité de Santé (HAS). Malgré cette non-recommandation, le nombre de dépistages individuels est élevé et conduit à la détection de formes précoces du CaP. Près de 40 à 50 % des cancers identifiés sont d'évolution lente et ne se seraient jamais révélés au cours de la vie des personnes.

2.5.1 Détection précoce

Toucher rectal Le toucher rectal (TR) est recommandé préalablement à la prescription du dosage du PSA total. Un TR suspect est une indication de biopsies prostatiques quelle que soit la valeur du PSA.

PSA Le PSA (*Prostatic Specific Antigen*) est une molécule qui n'est fabriquée que par la prostate. Le dosage du PSA dans le sang est un indicateur pour le diagnostic étant donné que le risque de CaP augmente avec la valeur du PSA total.

Une valeur supérieure à 4 ng/ml est généralement considérée comme anormale, mais cette valeur doit être interprétée par le médecin en fonction du contexte clinique. Il n'y a pas de valeur seuil du PSA en deçà de laquelle il n'y a aucun risque de cancer.

Toutefois, le PSA est caractéristique de l'épithélium prostatique et non du CaP ; il peut donc être élevé dans d'autres situations (infection de la prostate, adénome, hypertrophie bénigne, etc.). La cinétique du PSA (temps de doublement, vélocité) est particulièrement utile au suivi des patients après traitement.

La détection précoce du CaP repose également sur la recherche d'antécédents familiaux et de l'ethnie, en plus du toucher rectal et du dosage du PSA total.

2.5.2 Diagnostic

Imagerie

En cas de suspicion de CaP, la dernière version des *guidelines* urologiques recommande d'effectuer une acquisition IRM avant toute première série de biopsies prostatiques, contrairement à ce qui était fait auparavant [93]. En effet, l'IRM a fait ses preuves en permettant d'augmenter l'identification de CaP significatif et de guider les biopsies prostatiques sur ces lésions. L'IRM est composée de 3 séquences : pondérée en T2 (noté T2-w), de diffusion (noté DWI) ou dynamique, aussi dite de perfusion (noté DCE). Une présentation plus détaillée est décrite section 2.7.

Le score PI-RADS Le score PI-RADS (*Prostate Imaging–Reporting and Data System*) a été instauré dans le but de diminuer la variabilité inter-observateur et d'améliorer l'analyse des examens d'IRM prostatique pour diagnostiquer avec fiabilité les tumeurs significatives (de GS ≥ 7). Le score est obtenu en combinant l'analyse des séquences T2-w, DWI et DCE pour chacune des zones suspectes. Les critères diffèrent selon les zones (ZP ou ZT) et une séquence sera privilégiée en cas de séquences discordantes : le DWI pour la ZP et le T2-w pour la ZT (voir figure 2.3).

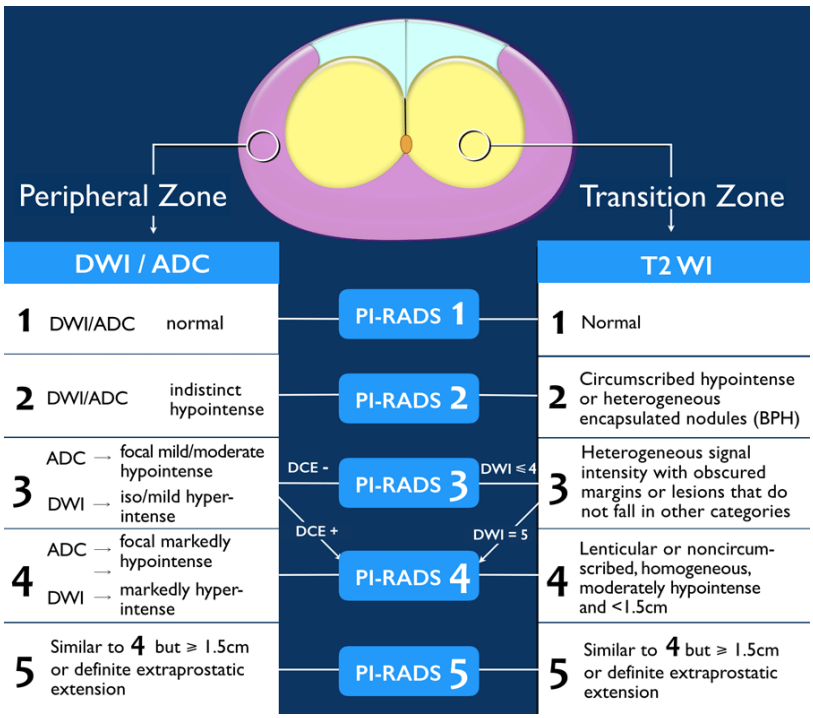


FIGURE 2.3 – Critère pour la notation des séquences T2-w et DWI/ADC en fonction de la zone considérée (ZP ou ZT) pour l'établissement du score PI-RADS final. DCE + correspond à un rehaussement focal précoce au niveau des zones suspectes. Source : de Rhiannon van Loenhout, Frank Zijta, Robin Smithuis et Ivo Schoots, Radiology assistant

Le PI-RADS varie de 1 à 5, correspondant à la probabilité que le cancer soit cliniquement significatif (CS) :

— PI-RADS 1 : très faible risque de cancer CS

— PI-RADS 2 : faible risque de cancer CS

— PI-RADS 3 : risque équivoque de cancer CS

— PI-RADS 4 : risque élevé de cancer CS

— PI-RADS 5 : risque très élevé de cancer CS

Des exemples de cas correspondants aux différents PI-RADS sont illustrées figure 2.4.

La stratification du risque tumoral en IRM repose sur le score PI-RADS, qui détermine en partie la stratégie de biopsie : simples biopsies non guidées par l'image (score PI-RADS <3), ou ajout de biopsies guidées par l'IRM si celle-ci est positive (score PI-RADS $\geq$3)

Biopsie

Une biopsie est un prélèvement d'une partie de tissu (ici de la glande prostatique) pour analyse. Il s'agit d'un examen invasif qui peut être douloureux et avoir des complications. Le schéma standard recommandé des biopsies systématiques correspond à 12 prélèvements (figure 2.5B). Pour chaque lobe, les prélèvements sont réalisés aux niveaux médial et latéral : à la base, au milieu et à l'apex. Il est important de noter qu'une biopsie est un échantillon de tissu localisé qui ne reflète qu'une zone de la prostate.

En cas d'IRM positive, des biopsies ciblées sont associées à des biopsies systématiques. L'échographie est l'examen de référence pour la réalisation des biopsies ciblées sur les lésions suspectes détectées à l'IRM, soit par un guidage visuel (repérage cognitif), soit par des techniques de fusion d'images IRM-échographie (figure 2.5A). Les biopsies ciblées sous IRM sont techniquement plus difficiles et coûteuses.

L'examen anatomopathologique (analyse au microscope des prélèvements tissulaires) permet de définir le score de Gleason et le groupe pronostique de la classification ISUP 2016 correspondant.

En cas de suspicion persistante de CaP après une première série de biopsies négatives, une deuxième série de biopsies prostatiques peut être indiquée. Il n'y a pas de consensus quant au meilleur délai entre les séries de biopsies. Avant une deuxième série de biopsies, les biopsies guidées par l'IRM augmentent de façon significative le taux de cancers cliniquement significatifs (ISUP grade $\geq$ 2).

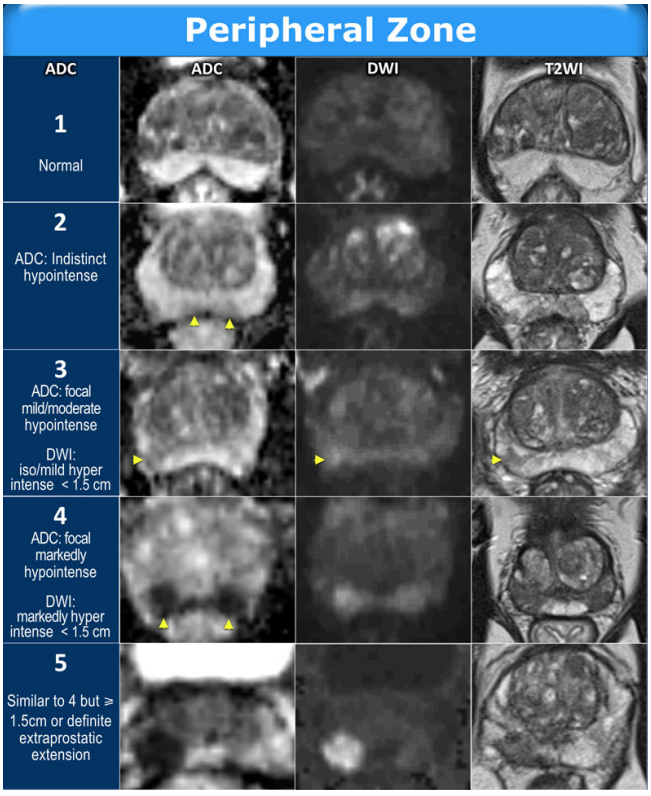
2.6 Traitement du cancer de la prostate

La prostatectomie totale (PT)

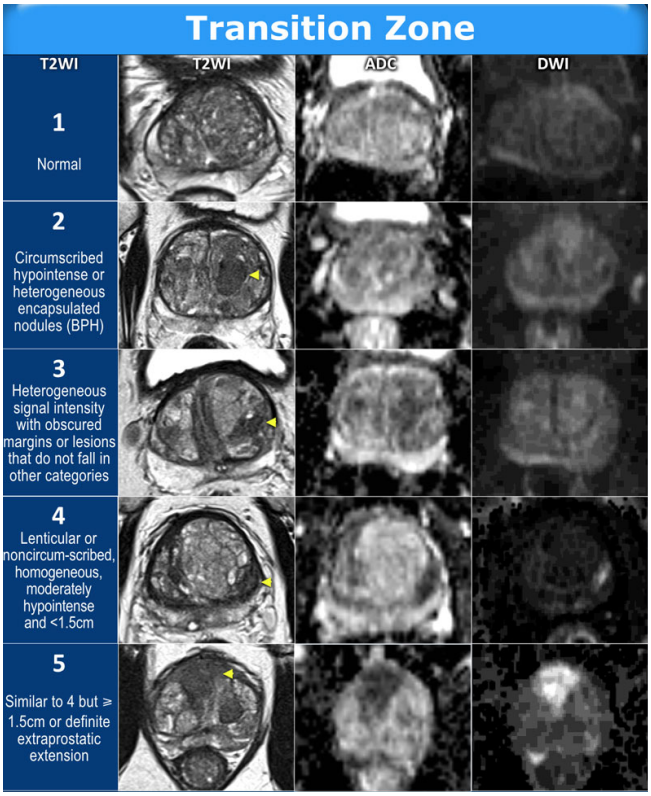
La PT est l'un des traitements de référence du CaP localisé, qui offre le plus de garantie carcinologique à long terme. L'objectif de la PT est l'ablation de la totalité de la prostate et des vésicules séminales. Cependant, ses effets secondaires inéluctables sur la continence urinaire et la fonction érectile - malgré la préservation sphinctérienne et des bandelettes neurovasculaires - ont un impact significatif sur la qualité de vie des patients traités.

La radiothérapie (RT)

La radiothérapie externe et la curiethérapie font partie des modalités thérapeutiques potentiellement curatives proposées dans le traitement du cancer de la prostate non métastatique.

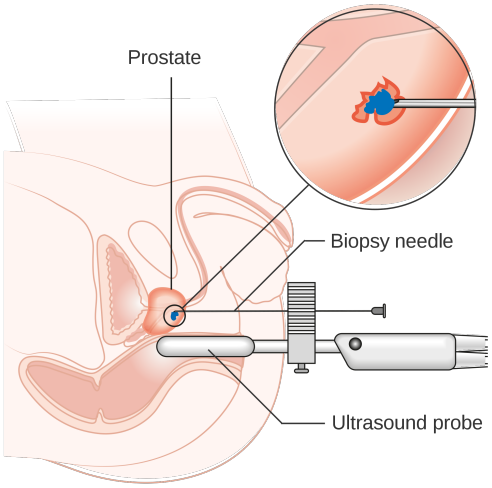


(A) Pour la ZP

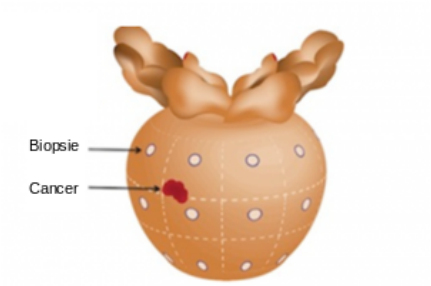


(B) Pour la ZT

FIGURE 2.4 – Exemple de cas correspondant aux différents scores PI-RADS pour les deux principales zones de la prostate. Les lésions sont identifiées par des flèches jaunes. Source : de Rhiannon van Loenhout, Frank Zijta, Robin Smithuis et Ivo Schoots, Radiology assistant



(A) Source : *Cancer Research UK*, via Wikimedia Commons



(B) Adaptée de CCM urology

FIGURE 2.5 – (A) Biopsie prostatique sous contrôle échographique. (B) Les 12 points de biopsie réalisés lors de biopsies aléatoires, pouvant passer à côté d'un cancer.

La curiethérapie ou radiothérapie interne

La curiethérapie est un traitement très localisé du cancer qui consiste à mettre en place des implants radioactifs (grains d'iode 125 ou sources d'iridium 192) à l'intérieur de la prostate. Ces implants émettent des rayonnements qui détruisent les cellules cancéreuses de la prostate. La curiethérapie est une modalité thérapeutique possible pour certains cancers de la prostate localisés à faible risque.

La radiothérapie externe (RTE)

La radiothérapie externe est un traitement local du cancer qui utilise des rayonnements ionisants qui détruisent les cellules cancéreuses en les empêchant de se multiplier. Elle consiste à diriger précisément ces rayonnements (appelés aussi rayons ou radiations) sur la zone à traiter, tout en préservant le mieux possible les tissus sains et les organes avoisinants, dits organes à risque (notamment la vessie et le dernier segment de l'appareil digestif : rectum et canal anal). Ces rayonnements sont produits par un appareil appelé accélérateur de particules. Ils sont dirigés en faisceaux vers la prostate pour atteindre, à travers la peau, la tumeur ainsi que les ganglions voisins. C'est également un traitement de référence, en association avec l'hormonothérapie, du CaP à haut risque et localement avancé. Elle s'accompagne également d'un risque d'impuissance et éventuellement d'incontinence urinaire, ainsi que d'un risque d'inflammation rectale (rectite radique).

L'hormonothérapie

Le CaP est un cancer dit hormonosensible, c'est-à-dire que son développement est stimulé par des hormones masculines : les androgènes et plus particulièrement la testostérone. L'hormonothérapie consiste à empêcher l'action stimulante de la testostérone sur les cellules cancéreuses pour stopper le développement du cancer. Son effet n'est que transitoire. Une hormonothérapie, associée à une RTE, est le traitement

de référence des cancers de la prostate localement avancés et un des traitements possibles des formes localisées à haut risque. L'hormonothérapie est habituellement débutée avant la radiothérapie puis poursuivie après la radiothérapie jusqu'à trois ans.

La chimiothérapie

La chimiothérapie est un traitement dont l'action est dirigée notamment sur les mécanismes de la division cellulaire. C'est un traitement général, dit aussi traitement systémique, qui agit dans l'ensemble du corps. Cela permet d'atteindre les cellules cancéreuses quelle que soit leur localisation. Elle est indiquée pour le traitement des cancers métastatiques hormonorésistants dans le but de soulager la douleur ou maîtriser les symptômes de la maladie.

Le HIFU

Le traitement par Ultrasons focalisés de haute intensité (*High-Intensity Focused Ultrasound*) - noté HIFU - est une thérapeutique non chirurgicale développée depuis une trentaine d'années pour des patients sélectionnés présentant un CaP localisé. C'est un traitement peu invasif qui traite le cancer de la prostate en concentrant des ultrasons focalisés de haute intensité qui vont détruire les cellules de la glande par la chaleur sans endommager les tissus environnants. Ce traitement local à la précision millimétrique permet de diminuer le risque d'effets secondaires. À noter que les premiers essais cliniques ont été initiés en 1993 à l'hôpital Edouard Herriot (Lyon, France), utilisant alors un prototype développé au laboratoire LabTAU (Lyon, France). Le projet RHU PERFUSE, financeur de cette thèse, s'inscrit dans la continuité de ces travaux et a pour but d'évaluer les résultats oncologiques du traitement focal HIFU.

Traitements focaux D'autres types de traitements focaux alternatifs sont étudiés mais encore en cours d'évaluation : la cryothérapie et le laser. De manière générale, la thérapie focale doit être considérée comme une technique en cours d'évaluation. Le traitement focal implique de connaître précisément la position des foyers tumoraux dans la glande.

La surveillance active

La surveillance active permet de différer la mise en route d'un traitement curatif (et des effets indésirables qui l'accompagnent). C'est une option thérapeutique de référence pour les tumeurs de faible risque évolutif. Le principe de la surveillance active repose sur des examens réguliers : examen clinique, dosages répétés du PSA pour suivre son évolution, biopsies prostatiques et éventuellement IRM. Si une évolution de la maladie est détectée, un traitement ayant pour objectif de traiter la maladie peut être programmé.

2.7 Imagerie du cancer de la prostate

Récemment, l'IRM s'est imposée comme technique d'imagerie pour le CaP. Elle permet de guider les biopsies, d'évaluer de manière non invasive le stade du cancer et de localiser le cancer pour un traitement focal.

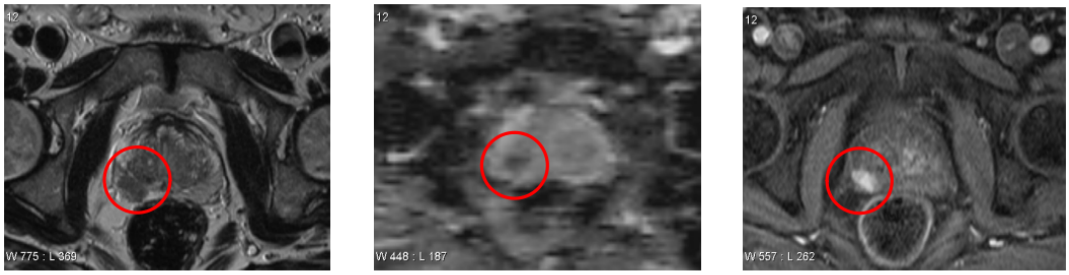


FIGURE 2.6 – IRM multiparamétrique de prostate. Ici, les coupes sont axiales dans le plan médian. La lésion visible est entourée en rouge sur les trois modalités d'imagerie.

L'IRM prostatique est un examen standardisé, qui doit suivre certains critères dans sa mise en place et son analyse.

2.7.1 Acquisition

L'IRM doit être multiparamétrique et composée des séquences :

- anatomiques
  - morphologiques (T2-w)
- fonctionnelles
  - de diffusion à haute valeur de b $\geq$ 1400 (DWI)
  - de perfusion T1 avec injection d'un produit de contraste (DCE)

Le plan de référence est le plan axial oblique perpendiculaire à la paroi rectale. L'IRM prostatique peut être effectuée à 1.5T ou à 3T. L'examen peut être effectué avec une antenne endorectale seule, avec une antenne externe (voir figure 2.7) ou en couplant ces deux antennes. Un examen complet dure de 20 à 30 minutes.

(A) Antenne pelvienne (B) Antenne endorectale (C) IRM Siemens 1.5T

FIGURE 2.7 – Antenne pelvienne (A), endorectale (B) et imageur IRM Siemens Magnetom Symphony 1.5T (C), utilisés pour certaines des acquisitions IRM-mp utilisées dans cette thèse. Adaptée de Niaf [94].

2.7.2 Imagerie en pondération T2 (T2-w)

La séquence pondérée en T2 est également désignée sous le terme de séquence "morphologique" : elle permet de visualiser les différentes zones de la prostate (ZP, ZT et ZC). C'est la séquence de référence pour visualiser les tissus prostatiques. Sur le T2-w, les lésions cancéreuses apparaissent en hyposignal.

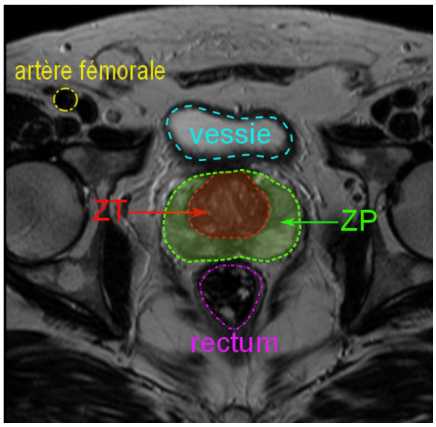


FIGURE 2.8 – IRM du pelvis masculin : séquence T2-w. Adaptée de Niaf [94]

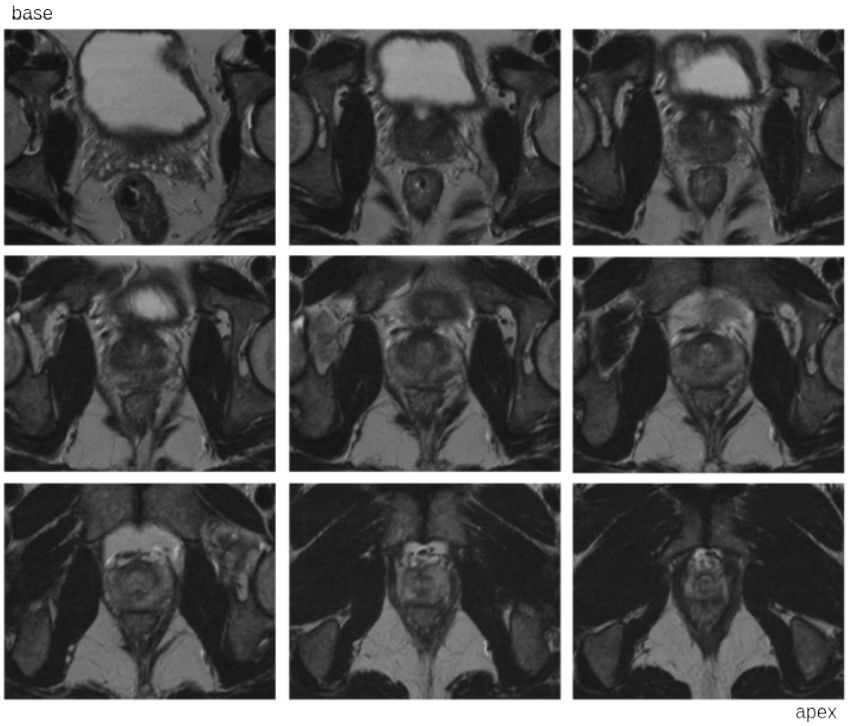


FIGURE 2.9 – Coupes axiales de la prostate en IRM T2-w acquises chez un patient de 53 ans. Deux lésions sont présentes en ZP postérieure droite et gauche. Source : Niaf [94]

2.7.3 Imagerie de diffusion (DWI)

Les séquences de diffusion permettent l'étude des mouvements moléculaires de l'eau. L'acquisition consiste à appliquer à une séquence rapide d'IRM plusieurs gradients de diffusion b. Plus la valeur de b est élevée, plus le signal décroît. Dans les tissus anormaux, le signal se maintient malgré l'augmentation de b et les tumeurs apparaissent en hyper-signal sur la diffusion. La pente de la courbe du signal en fonction des valeurs de b permet d'extraire une carte de caractéristiques appelée l'ADC (le coefficient apparent de diffusion, de l'anglais *Apparent Diffusion Coefficient*), sur laquelle les cancers apparaissent en hyposignal.

La diffusion donnerait aussi des informations sur l'agressivité tumorale. Les

études ne rapportent pas toutes les mêmes conclusions, mais il semblerait qu'il existe une corrélation inversement proportionnelle entre la valeur de l'ADC et le score de Gleason : plus l'ADC est bas, plus le score de Gleason serait élevé. Une récente revue de littérature de 39 études a conclu sur une corrélation modérée dans la ZP et faible dans la ZT [120].

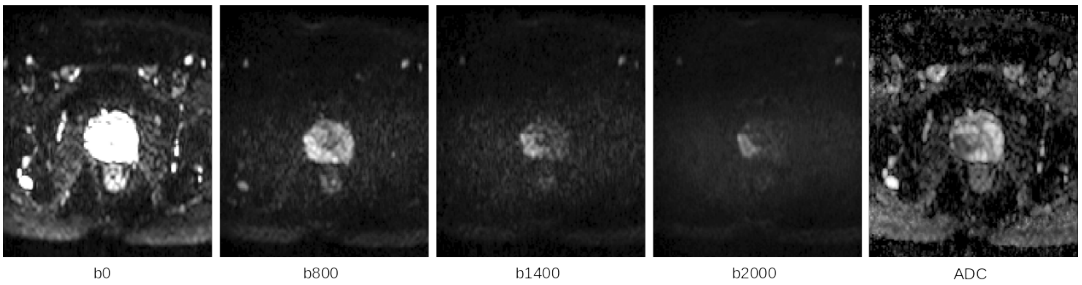


FIGURE 2.10 – Coupes axiales en IRM de diffusion pour différentes valeurs de *b* et carte ADC correspondante. Une lésion GS 9 en ZP est visible en hyper-signal sur les séquences DWI à haute valeur de b et en hyposignal sur l'ADC.

2.7.4 Imagerie de perfusion (DCE)

L'IRM de perfusion permet l'étude de la microvascularisation (les tissus atteints sont généralement riches en micro vaisseaux sanguins). L'acquisition repose sur l'injection intraveineuse en bolus d'un d'agent de contraste (gadolinium), dont on peut suivre le signal en fonction du temps. Pour cela, de courtes séquences d'écho de gradient T1 de 10-15 secondes sont répétées sur 2-3 min. La séquence est également appelée séquence "dynamique".

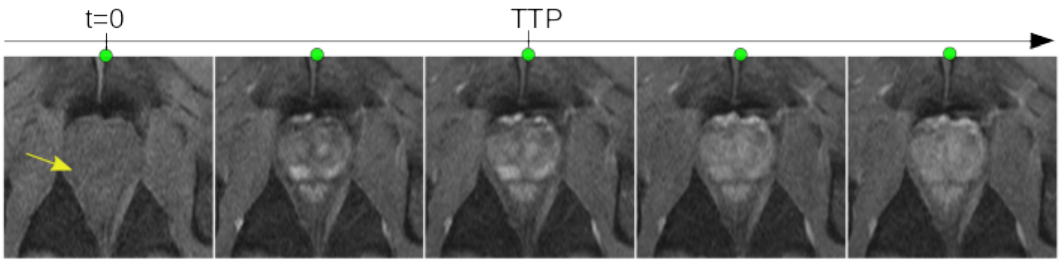


FIGURE 2.11 – Images correspondant à différents temps d'une séquence IRM de perfusion. L'agent de contraste est injecté à t=0 et le temps au pic (noté TTP, de l'anglais *Time To Peak*) correspond à la 3ème image. La région maligne est indiquée par la flèche jaune. Adaptée de Niaf [94].

L'analyse de la courbe de l'intensité dans le temps permet d'extraire des paramètres discriminants tels que :

— le pic de rehaussement maximal (ou *maximum enhancement)* : l'intensité maximale du signal au cours du temps

— le temps au pic (ou TTP pour *time to peak*) : le temps entre le début du rehaussement et le pic de rehaussement

— la vitesse de rehaussement (ou *wash-in*) : la pente de la droite entre le début du rehaussement et le maximum de rehaussement

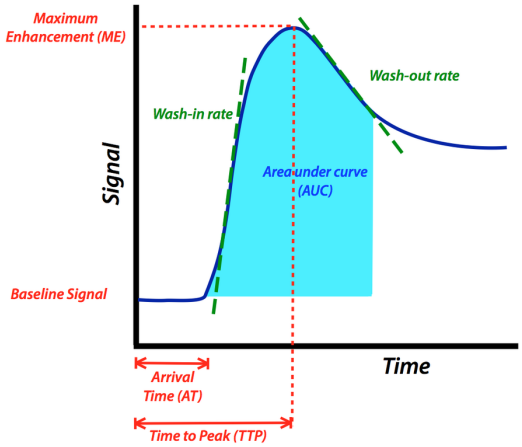


FIGURE 2.12 – Extraction de paramètres quantitatifs de la courbe présentant le signal moyen en fonction du temps de la séquence DCE. Source : MRIquestions.com, avec l'autorisation d'Allen D. Elster.

— la vitesse de lavage (ou *wash-out*) : la pente de la droite de régression linéaire à la courbe de rehaussement entre le maximum de rehaussement et la fin de la séquence

— l'aire sous la courbe de rehaussement (AUC) : l'intégrale de la courbe entre le début du rehaussement et la fin de la séquence

Un cancer est caractérisé par une prise de contraste précoce, intense et avec un lavage rapide (*wash-out*).

Un modèle pharmacocinétique (bicompartimental de Tofts) peut également être établi pour calculer notamment la constante de transfert K^{trans}, qui traduit le transfert du produit de contraste vers l'interstitium (*wash-in*).

2.7.5 Limites de l'IRM pour le diagnostic du CaP

L'IRM a fait ses preuves ces dernières années et est dorénavant au cœur du diagnostic et du suivi des patients pour le CaP. L'analyse jointe des séquences morphologiques et fonctionnelles améliore la détection des cancers.

Toutefois, l'analyse jointe des séquences IRM est une tâche chronophage et qui peut être rendue difficile notamment dans les cas suivants :

— présence d'artefacts dus à des foyers hémorragiques post biopsiques

— prostatite (inflammation de la prostate)

— présence de nodules (formations anormales de forme ronde) bénins, pouvant être pris pour des lésions

— séquences discordantes : la lésion peut n'être visible que sur une ou deux séquences IRM

— cancers non visibles à l'IRM (de faibles grades la plupart du temps)

Pour toutes ces raisons, la variabilité intra et inter-observateur est élevée [54, 115], des radiologues expérimentés obtiennent une meilleure sensibilité lors de l'analyse de séquences IRM-mp.

Pour pallier ces difficultés, des directives pour l'analyse de séquences et l'établissement de critères tels que PI-RADS sont mis en place et sont en perpétuelle amélioration.

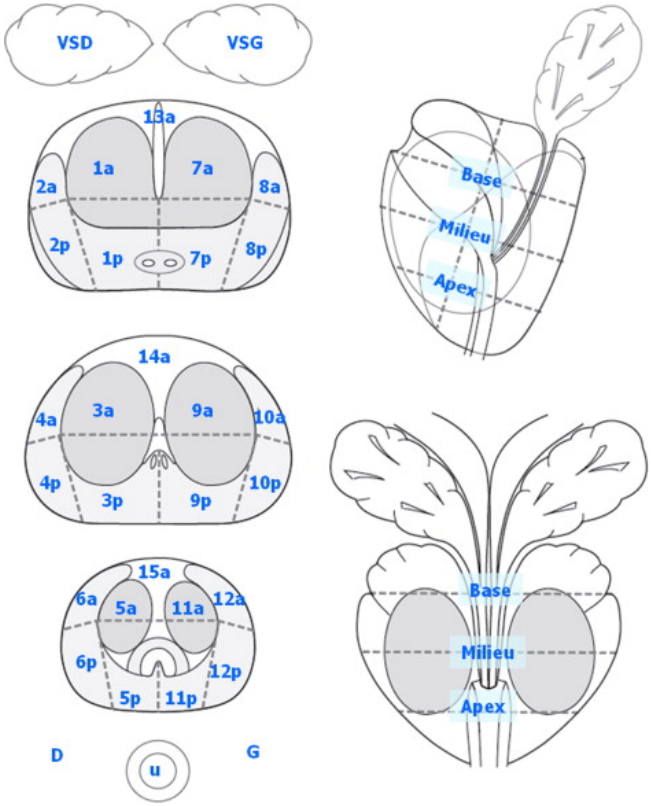


FIGURE 2.13 – Cartographie de la prostate en 27 secteurs utilisée pour localiser les lésions suspectes vues à l'IRM. Les noms de secteurs finissant par un *p* font partie de la ZP alors que ceux finissant par un *a* sont dans la ZT. Figure 2 de Puech et al. [100].

Malgré tous ces efforts, l'analyse reste une tâche difficile et fastidieuse. En outre, certaines informations ne sont actuellement pas décelables sur les images, telles que l'agressivité de la lésion (ou score de Gleason). Cette information serait intéressante en clinique, afin d'éviter ou de limiter les biopsies en ciblant les zones suspectes, ou de permettre une surveillance active non invasive des patients présentant des cancers peu agressifs.

C'est pourquoi de nombreuses recherches sur les systèmes d'aide au diagnostic (CAD) sont en cours pour assister le radiologue dans l'analyse des images IRM-mp.

2.8 Conclusion

Le cancer de la prostate (CaP) est un problème de santé publique majeur. Son diagnostic est en perpétuelle évolution, avec la récente inclusion de l'IRM dans le processus de diagnostic avant même la réalisation de biopsies, suite à la démonstration de l'augmentation des performances en ciblant les biopsies.

Toutefois, l'analyse jointe de ces images multimodales est fastidieuse et chronophage, en particulier lorsque les séquences mènent à des conclusions différentes. En outre, la sensibilité de l'IRM reste faible pour les cancers de petit score de Gleason, et la variabilité inter-observateur élevée.

C'est pourquoi de nombreux systèmes d'aide au diagnostic (CAD) ont été proposés ces dernières années [131].

Dans le chapitre suivant, nous présentons les bases de l'apprentissage profond pour la segmentation d'images médicales et l'établissement de systèmes d'aide au diagnostic.

Chapitre 3

L'apprentissage profond en segmentation d'images médicales

Sommaire

3.1 Introduction

Dans ce chapitre, il s'agira de rappeler les bases des algorithmes d'apprentissage nécessaires à la compréhension de la suite du manuscrit, en mettant l'accent sur les réseaux de segmentation d'images. Pour un cours complet sur les réseaux de neurones appliqués à la vision par ordinateur, le cours CS231n de l'Université de Stanford est très approprié avec vidéos, supports de cours et travaux pratiques disponibles gratuitement en ligne. Pour un livre intégral sur l'apprentissage automatique, les livres de Bishop [8] ou de Goodfellow et al. [52] sont des références dans le domaine.

3.2 Apprentissage automatique

3.2.1 Généralités

L'apprentissage automatique (ou *machine learning*) est une sous-classe de l'intelligence artificielle (IA) permettant à un système d'apprendre à partir de **données** et non d'une suite d'opérations séquentielles déterminées à l'avance.

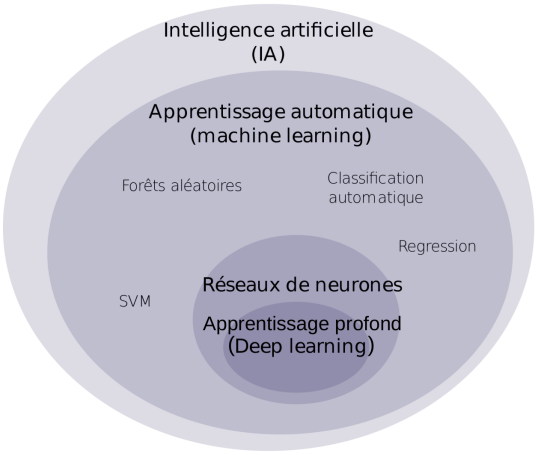


FIGURE 3.1 – Relation entre l'IA, l'apprentissage automatique, les réseaux de neurones et l'apprentissage profond. Adaptée de Wikipedia.

Soit un ensemble d'entraînement constitué de N paires de données annotées $D = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_N}, y_N)\}$ avec $\mathbf{x_i} \in \mathbb{R}^d$ une donnée (représentée par un vecteur) et y_i la cible associée à $\mathbf{x_i}$. L'objectif est d'apprendre une fonction de décision qui sache prédire y_i à partir de $\mathbf{x_i}$ telle que :

$$f : \mathbf{x_i} \rightarrow y_i. \tag{3.1}$$

Une autre définition possible est celle de Mitchell [92] : "On dit d'un programme informatique qu'il apprend de l'**expérience E** en ce qui concerne une certaine **tâche T** et une mesure de **performance P**, si sa performance à la tâche T, telle que mesurée par P, s'améliore avec l'expérience E."

Cette définition introduit trois notions clés de l'apprentissage automatique, détaillées ci-après.

La **tâche T** est l'action que le programme informatique va réaliser. Il peut s'agir de traduire un discours, estimer une densité de probabilité, participer à la conduite d'un véhicule autonome ou reconnaître la présence d'un organe dans une image. Ces différentes tâches peuvent être réparties en plusieurs grandes catégories, parmi lesquelles la **classification**. Le but d'un algorithme de classification est d'assigner une étiquette à une observation donnée, tel que $f : \mathbb{R}^d \rightarrow \mathbb{N}$. Par exemple, $\mathbf{x_i}$ peut être une image de scanner des poumons, et les différentes classes des maladies associées qu'il s'agira d'identifier (classe 0 : patient sain, classe 1 : pneumopathie, classe 2 : cancer). Une autre tâche est la **régression**, où la variable cible est continue. La fonction sera alors de type $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Un exemple de tâche de régression serait de prédire l'âge d'un patient à partir d'un scanner des poumons. La tâche étudiée dans cette thèse est celle de la **segmentation**, qui correspond à une tâche de classification sur chacun des pixels d'une image en entrée. La fonction apprise est $f : \mathbb{R}^d \rightarrow \mathbb{N}^d$. Visuellement, cela correspond à délimiter (ou contourer) des objets d'intérêt. Toujours avec un scanner des poumons en exemple, une tâche de segmentation serait de prédire pour chaque pixel s'il correspond à du tissu sain ou lésionnel.

La mesure de **performance P** sert à évaluer le modèle f appris. Selon la tâche étudiée, différentes métriques sont utilisées (détaillées section 3.5.2). Les performances sont mesurées à partir de l'ensemble de données D utilisé pour l'apprentissage mais

également sur de nouvelles données pour vérifier que le modèle **généralise** correctement. Pour ce faire, seule une partie des données disponibles est utilisée pour l'entraînement, l'autre étant réservée pour une phase de **test**.

Enfin, l'**expérience E** fait référence à la source des exemples donnés pour l'apprentissage. Dans cette thèse, l'expérience correspond aux ensembles de données disponibles, c'est-à-dire les images d'IRM de prostate accompagnées de leurs annotations.

La fonction f est définie par des paramètres W, qui sont appris à partir des données. Pour trouver les paramètres W^* optimaux pour la fonction f, on cherche à minimiser une **fonction de coût** ou **risque statistique** :

$$W^* = \arg\min_W R_{\text{réel}}\big(f_W(\mathbf{x})\big) \tag{3.2}$$

avec

$$R_{\text{réel}}\big(f_W(\mathbf{x})\big) = \mathbb{E}\big[L\big(f_W(\mathbf{x}), y\big)\big] = \int_{X\times Y} L\big(f_W(\mathbf{x}), y\big) dP(\mathbf{x}, y) \tag{3.3}$$

où $P(\mathbf{x}, y)$ est la vraie distribution de probabilités d'où sont issus $\{\mathbf{x_i}, y_i\}_{1\leq i\leq N}$. En pratique, $P(\mathbf{x}, y)$ n'est pas connue et doit être approximée par le **risque empirique** (via une approximation Monte Carlo) sur le jeu de données d'entraînement D_{train}, défini par :

$$R_{\text{emp}}\big(f_W(\mathbf{x})\big) = \frac{1}{N} \sum_{\mathbf{x},y\in D_{train}} L\big(f_W(\mathbf{x}), y\big). \tag{3.4}$$

Un terme supplémentaire est ajouté à ce risque empirique pour imposer une distribution *a priori* sur les paramètres du modèle. Ce terme est généralement sous la forme d'une pénalité envers la complexité du modèle f et est appelé **régularisation**. La fonction de coût à minimiser s'écrit alors :

$$R\big(f_W(\mathbf{x})\big) = R_{\text{emp}}\big(f_W(\mathbf{x})\big) + \lambda R_{\text{struct}}(f) \tag{3.5}$$

où λ est un paramètre à définir pour choisir la force de la régularisation.

Le modèle est également défini par un certain nombre d'**hyperparamètres** (dont fait d'ailleurs partie la force de la régularisation) : ce sont des paramètres qui ne peuvent être appris à partir des données et doivent être fixés manuellement par un utilisateur. Par exemple, on peut citer le nombre d'arbres dans une forêt aléatoire, le nombre de voisins dans la méthode des k plus proches voisins ou encore le taux d'apprentissage dans un réseau de neurones.

3.2.2 Familles d'algorithmes

Plusieurs types de modèles statistiques, linéaires ou non linéaires, peuvent être utilisés selon la tâche, la nature et le volume des données à traiter. Les modèles statistiques d'apprentissage automatique sont définis par différents éléments, parmi lesquels la fonction de décision f, le type de données d'apprentissage à disposition et la méthode d'optimisation requise pour minimiser la fonction de coût.

Fonction de décision Le modèle statistique sous-jacent représente une contrainte sur le type de fonction de décision f, non dépendante des données. Par exemple, dans le cas d'une régression linéaire, le modèle est contraint à avoir la forme d'une

droite en 2D, d'un plan en 3D et d'un hyperplan en K-D où $K > 3$. Il est donc important d'adapter le modèle à la tâche et aux données disponibles. Le but est de trouver un modèle qui représente suffisamment les données sans être trop spécifique à ces dernières. Un modèle trop complexe au regard des données d'entraînement disponibles risque de **sur-apprendre** (*overfitting*) alors qu'un modèle sous-dimensionné pourra ne pas avoir un pouvoir représentatif suffisant et ainsi **sous-apprendre** (voir figure 3.2).

Parmi les modèles classiques, on peut citer des modèles linéaires ou non linéaires, tels que les modèles de régression linéaire ou logistique, l'analyse discriminante linéaire, le perceptron, les séparateurs à vaste marge (SVM), les arbres de décision, les méthodes des k plus proches voisins, ou encore les réseaux de neurones.

Données d'apprentissage Selon les informations disponibles durant la phase d'apprentissage, le modèle est qualifié de différentes manières :

- *supervisé* si les données sont étiquetées, c'est-à-dire que y_i ou la cible est connue pour les données utilisées en entraînement ;
- *par renforcement* quand le modèle est appris de manière incrémentale en fonction d'une récompense retournée par l'environnement pour chacune des actions entreprises ;
- *non supervisé* si les données n'ont pas d'étiquette. On cherche alors à déterminer la structure sous-jacente des données ou les lois $p(x)$, $p(x, y)$ et $p(x|y)$ suivant généralement des approches bayésiennes et diverses hypothèses quant à la distribution et l'indépendance des données ;
- *semi-supervisé* si une partie des données est étiquetée et une autre partie ne l'est pas (au croisement de l'apprentissage supervisé et non supervisé) ;
- *faiblement supervisé* si les étiquettes des données sont partiellement connues ou sont bruitées. Ce type d'approche sera détaillé dans le chapitre 8.

Méthode d'optimisation Les modèles se distinguent également par la méthode d'optimisation de la fonction de coût utilisée. Elle peut se faire en "*close form*", itérative ou récursive (c'est le cas des arbres de décision). De plus, le choix de la fonction de coût à optimiser va être caractéristique du modèle.

Par la suite, nous nous focaliserons sur les réseaux de neurones à la base des modèles par apprentissage profond.

3.2.3 Le perceptron multicouches

L'exemple le plus simple d'un réseau neuronal est le perceptron, proposé à la fin des années 50 [103] par analogie simpliste avec un neurone biologique (voir figure 3.3). Un neurone est représenté par des poids w_i, qui vont interagir de manière multiplicative avec les signaux x_i reçus des autres neurones. Le signal résultant est transmis à d'autres neurones, après une non-linéarité g.

Le perceptron correspond à la formule mathématique suivante :

$$f_W(\mathbf{x}) = g\left(\sum_i w_i x_i + w_0\right) \tag{3.6}$$

avec g une fonction d'activation non-linéaire. Les paramètres w_i sont déterminés itérativement lors de l'apprentissage du modèle.

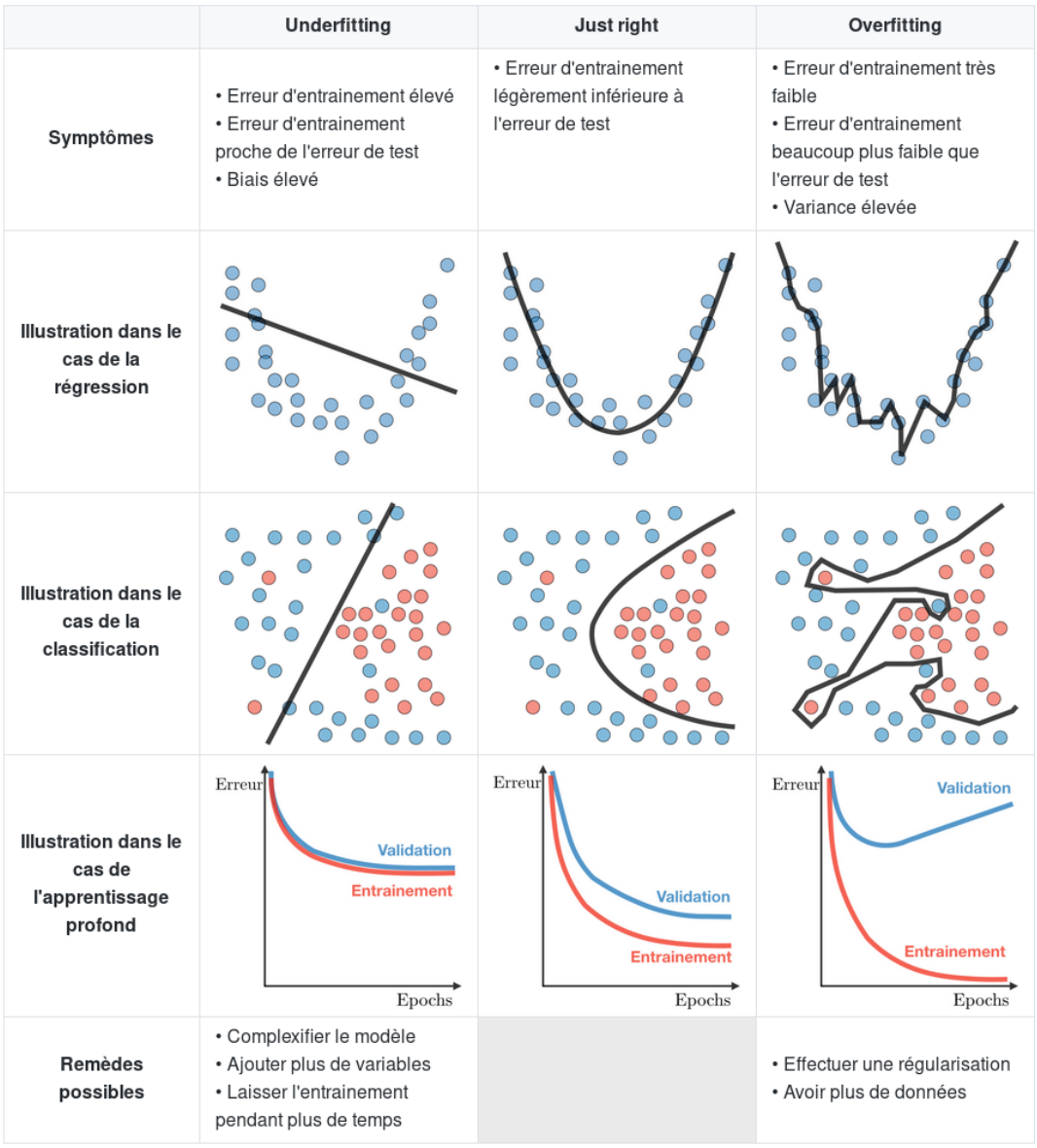


FIGURE 3.2 – Établissement d'un modèle : compromis à faire dans la complexité du modèle. Un modèle trop simple ne vas pas correctement représenter les données (sous-apprentissage) et un modèle trop complexe va être trop spécifique aux données (sur-apprentissage). Source : CS229, avec l'autorisation de Afshine Amidi et Shervine Amidi, 2018

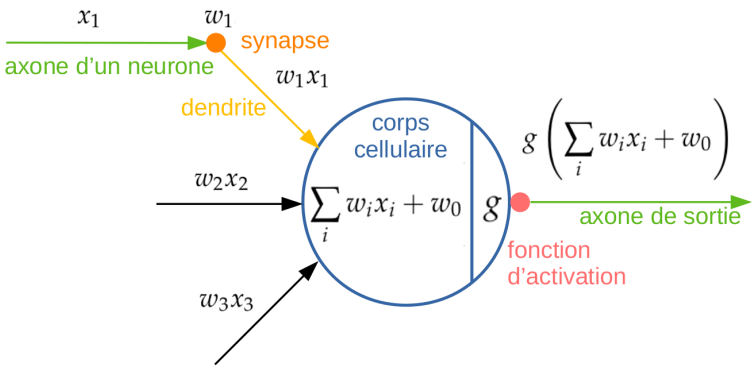


FIGURE 3.3 – Modèle mathématique d'un neurone. Inspirée de CS231n.

Reprenons maintenant l'équation 3.6 pour un problème à K dimensions et 2 classes :

$$\begin{aligned} f_W(\mathbf{x}) &= g(\sum_i x_i w_i + w_0) \\ &= g(w_0 + w_1x_1 + w_2x_2 + ... + w_Kx_K) \\ &= g\big((w_0, w_1, w_2, ..., w_K).(1, x_1, x_2, ..., x_K)\big) \\ &= g(W^T.\mathbf{x}') \end{aligned}$$

avec $\mathbf{x}' = [1, \mathbf{x}]$ une transformation du vecteur de donnée permettant d'inclure le biais w_0 dans le vecteur de poids W. L'équation du perceptron correspond donc à l'équation d'un hyperplan utilisé pour un problème de classification linéaire à deux classes, si la non-linéarité g correspond à la fonction signe. Le biais est considéré comme inclus dans les vecteurs sans modifier les notations de $\mathbf{x}$ et W, pour des raisons de simplicité.

Ce modèle permet donc de résoudre des problèmes linéaires. Or, tous les problèmes ne sont pas linéairement séparables, ce qui est limitant. En 1986, le perceptron multicouche (voir figure 3.5) est proposé par David Rumelhart [107], permettant d'apprendre des modèles non-linéaires par l'utilisation de **fonctions d'activation** (ou non-linéarité) entre les différentes couches. Les fonctions d'activations les plus courantes sont présentées à la figure 3.4.

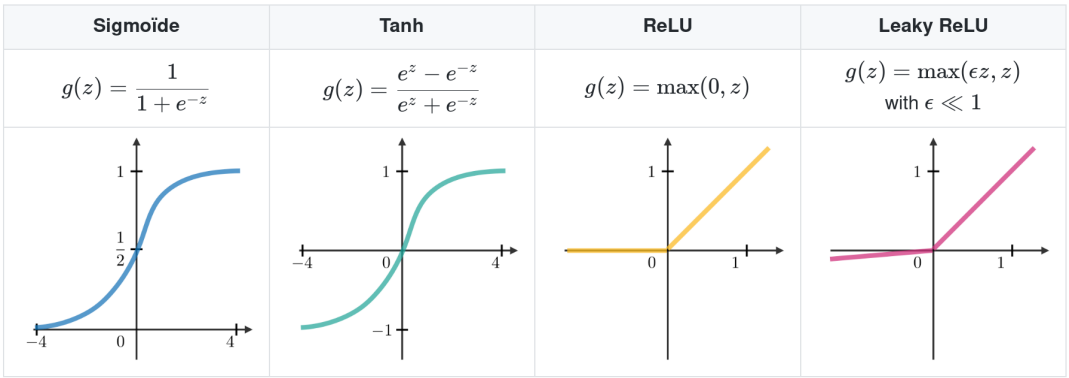


FIGURE 3.4 – Fonctions d'activation les plus courantes. Source : CS229, avec l'autorisation de Afshine Amidi et Shervine Amidi, 2018

Les perceptrons multicouches sont à la base des réseaux de neurones profonds.

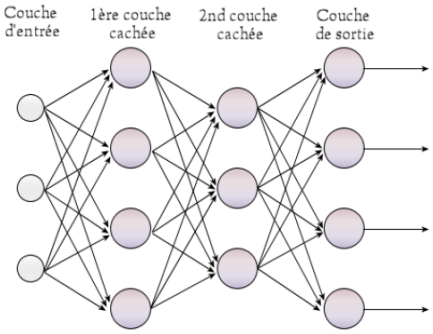


FIGURE 3.5 – Un perceptron multicouches. Source : Wikimedia Commons

En effet, un réseau de neurones profond est constitué d'un ensemble de neurones (ou perceptrons), connectés dans un graphe acyclique, souvent organisé par couches (figure 3.5).

C'est cette notion de réseaux multicouches qui est à la base de ce qu'on appelle l'**apprentissage profond**.

3.3 Les réseaux de neurones convolutionnels (CNN)

Dans des réseaux de neurones classiques, les couches les plus communes sont des couches totalement connectées (*fully-connected*) où les neurones entre deux couches adjacentes sont connectés deux à deux, mais des neurones au sein d'une même couche ne partagent pas de connexion (voir section 3.3.1.3). Les réseaux de neurones convolutionnels (CNNs) sont très similaires à des réseaux de neurones classiques. Leur particularité dérive d'opérations de convolution au lieu de couches totalement connectées, opérations s'appliquant à des signaux spatialement structurés comme du texte, des signaux audios, des images, dont ces réseaux exploitent au maximum les propriétés. Un CNN est composé de trois types de couches principales : les couches de convolution, de *pooling*, ainsi que des couches totalement connectées. La concaténation de toutes ces couches forme un CNN. Ces couches sont présentées plus en détail dans la partie suivante.

3.3.1 Couches composant les CNN

3.3.1.1 Couche de convolution

Une convolution 2D consiste à glisser un filtre sur une image pour en extraire des caractéristiques. En apprentissage profond, la couche de convolution correspond en pratique à l'opération mathématique de corrélation croisée. On utilisera par la suite le terme de convolution dans le sens de l'apprentissage profond.

La valeur du pixel (i, j) en sortie d'une convolution entre une image I et un filtre K se calcule dans un CNN de la façon suivante :

$$(I * K)(i, j) = \sum_m \sum_n I(m, n) K(i + m, j + n). \tag{3.7}$$

Le résultat d'une convolution est nommée carte de caractéristiques (*feature map*).

L'utilisation de couches de convolution permet de réduire considérablement le nombre de paramètres par rapport à des couches totalement connectées, puisqu'un

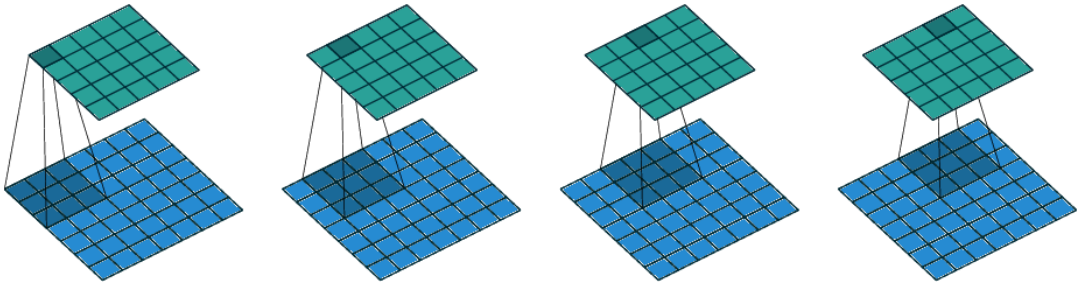


FIGURE 3.6 – Opération de convolution sur une image 7×7. Le même filtre de taille 3×3 se déplace sur l'intégralité de l'image avec un pas de 1, pour extraire une carte de caractéristiques de taille 5×5. Source : Vincent Dumoulin et Francesco Visin.

même filtre est appliqué sur toute l'image (voir figure 3.6). Ainsi, les poids sont **partagés** entre les différents neurones.

3.3.1.2 Couche de *pooling*

Des couches de *pooling* sont fréquemment insérées entre les couches de convolution dans un CNN. Cette couche permet de réduire la taille spatiale de la représentation afin de réduire le nombre de paramètres du réseau, et donc de limiter le sur-apprentissage. La couche de *pooling* opère indépendamment selon la dimension z sur chaque carte de caractéristiques en entrée et la redimensionne spatialement dans le plan (x, y), en utilisant une opération telle que le maximum ou la moyenne. La forme la plus courante correspond à un *max pooling* avec des filtres de taille 2×2 appliqués avec un pas de 2, rejetant 75 % des activations, sans modifier la dimension de la profondeur (voir figure 3.7).

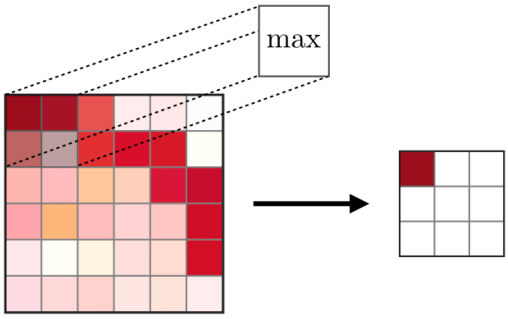


FIGURE 3.7 – Les couches de *pooling* permettent de sous-échantillonner le volume et réduire le nombre de paramètres. Ici, l'opération la plus courante (*max pooling*) est présentée : elle consiste à prendre la valeur maximale de chaque région. La valeur moyenne peut également être utilisée. Source : CS230, avec l'autorisation de Afshine Amidi et Shervine Amidi, 2018

3.3.1.3 Couche totalement connectée (*Fully Connected*, notée FC)

La couche FC s'applique sur une entrée où chaque neurone est connectée à tous les neurones. Les couches FC sont généralement présentes à la fin des architectures de CNN, avant le classifieur (présenté section 3.3.1.5).

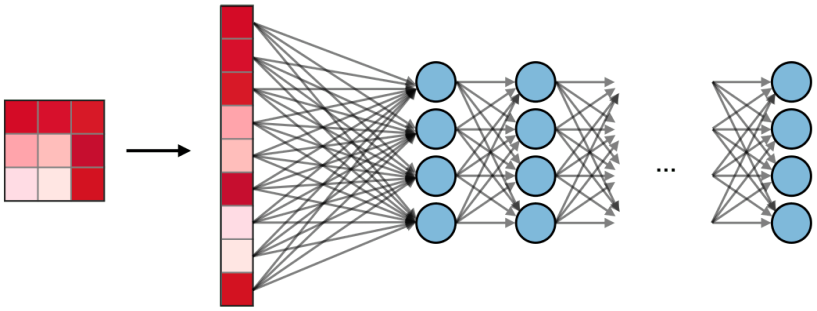


FIGURE 3.8 – Couche totalement connectée (FC), où chaque neurone d'une couche est connecté à tous les neurones de la couche suivante. Source : CS230, avec l'autorisation de Afshine Amidi et Shervine Amidi, 2018

3.3.1.4 Couche de normalisation

Les couches de normalisation, quand elles sont utilisées, s'insèrent après une couche de convolution et avant la fonction d'activation (voir figure 3.4). La normalisation permet d'utiliser un taux d'apprentissage plus grand, de limiter la disparition des gradients et de réduire une dépendance trop forte à l'initialisation. Elle est effectuée par *batch* de k éléments, que l'on note $x^{(k)}$:

$$x^{(k)} \leftarrow \gamma \frac{x^{(k)} - \mathbb{E}(x^{(k)})}{\sqrt{\text{Var}(x^{(k)}) + \epsilon}} + \beta \tag{3.8}$$

avec $\mathbb{E}$ l'espérance mathématique, Var la variance, γ et β deux paramètres appris par le réseau et ϵ une constante permettant d'éviter la division par zéro.

3.3.1.5 Couche de classification

La couche de classification correspond à la dernière couche du réseau. Généralement, on utilise la **sigmoïde** pour un problème binaire et le classifieur **softmax** pour un problème multiclasse. Le classifieur softmax se base sur la régression linéaire et se nomme régression logistique multinomiale :

$$P(Y = k | X = \mathbf{x_i}, W) = \frac{e^{s_k}}{\sum_j e^{s_j}} \tag{3.9}$$

avec $s = f_W(x_i)$, un vecteur de scores ou probabilités logarithmiques non normalisées (aussi appelées *logits*) pour chacune des classes k.

Avec l'exponentielle, les *logits* deviennent des probabilités (non normalisées), que la division normalise de sorte que leur somme égale 1. Cette expression peut être interprétée comme la probabilité (normalisée) que l'image $\mathbf{x_i}$ appartienne à la classe $Y = k$, selon les paramètres $\mathbf{W}$.

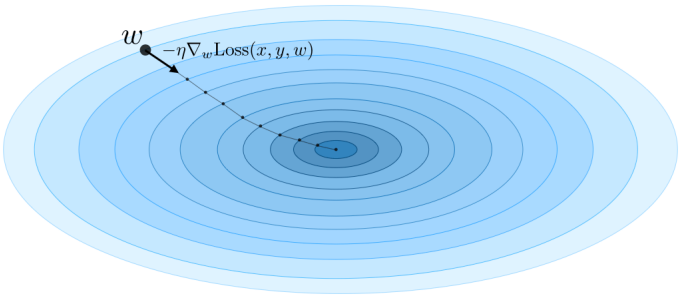


FIGURE 3.9 – Optimisation d'une fonction de coût par descente de gradient, conditionnée par le taux d'apprentissage (*learning rate*) η. Source : CS221, avec l'autorisation de Afshine Amidi et Shervine Amidi, 2018

3.3.2 Apprentissage des paramètres

Tel que mentionné à la section 3.2.1, l'apprentissage d'un modèle a pour but de trouver la valeur optimale de ses paramètres W. Pour apprendre ces paramètres, il s'agit de minimiser une fonction de coût L (*loss function* en anglais). En apprentissage profond, l'optimum d'une fonction ne peut être calculé directement de par sa complexité et sa forme non convexe. L'approche itérative par **descente de gradient** est utilisée dans ce cas. Toutefois, sachant que la fonction n'est pas convexe, l'optimisation peut ne pas mener à un minimum global mais à un minimum local.

L'apprentissage des paramètres d'un réseau de neurones se décompose en deux étapes : la propagation des données par l'avant (*forward step*) puis par l'arrière (*backward step*).

Propagation avant Étant donné une paire $(\mathbf{x_i}, y_i)$ correspondant à un exemple de la base d'apprentissage $\mathbf{x_i}$ et à sa vérité terrain (ou étiquette) y_i, la propagation avant consiste à donner l'exemple $\mathbf{x_i}$ en entrée du réseau et à le faire circuler dans toutes les couches cachées du réseau jusqu'à la couche de sortie, où un résultat $\hat{y}_i$ est obtenu. Une fonction de coût $L(\hat{y}_i, y_i)$ compare alors la prédiction à la vérité terrain pour mesurer l'erreur de prédiction.

Rétropropagation Le but étant de minimiser l'erreur de prédiction mesurée, les paramètres W du modèle sont mis à jour afin de minimiser cette erreur L. Pour cela, on cherche la direction dans laquelle L diminue le plus rapidement, c'est-à-dire où la pente est la plus importante. La direction du gradient correspond à celle pour laquelle l'accroissement de la fonction est maximale. Ainsi, les poids du modèle sont actualisés selon la direction opposée au gradient (le but étant de minimiser et non de maximiser l'erreur) par **descente de gradient** :

$$w \leftarrow w - \eta \frac{\partial L(\hat{y}, y)}{\partial w} \tag{3.10}$$

avec le scalaire η le taux d'apprentissage (*learning rate*), qui précise la taille du pas dans la direction opposée au gradient à chaque itération. Cet hyperparamètre est crucial dans l'entraînement d'un réseau de neurones et peut être optimisé selon différentes stratégies (présentées à la fin de cette section).

La rétropropagation du gradient (en anglais *backpropagation*) est une méthode destinée à calculer la dérivée partielle de la fonction de coût par rapport aux paramètres du modèle. Cette solution propose une évaluation analytique du gradient $\nabla L(\hat{y}, y)$ basée sur l'application du théorème de dérivation des fonctions composées ou règle de dérivation en chaîne (*chain rule*) :

$$\frac{\partial L(\hat{y}, y)}{\partial w} = \frac{\partial L(\hat{y}, y)}{\partial x} \times \frac{\partial x}{\partial y} \times \frac{\partial y}{\partial w}. \tag{3.11}$$

De cette manière, le gradient peut être décomposé couche par couche en commençant depuis la fonction de coût, puis en propageant aux couches précédentes (voir figure 3.10). Le produit des gradients des couches nous indique comment les paramètres du modèle doivent évoluer afin de minimiser l'erreur globale. L'étape de mise à jour des paramètres est déterminée par un algorithme d'optimisation numérique, tel que la descente de gradient stochastique (SGD) ou une de ses variantes.

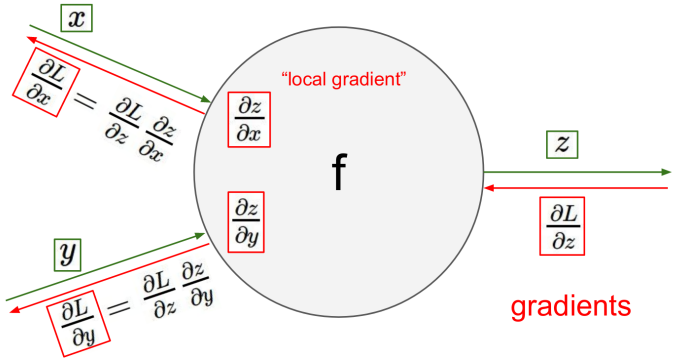


FIGURE 3.10 – Rétropropagation du gradient par dérivation des fonctions composées. La propagation avant correspond aux flèches vertes et la propagation arrière aux rouges. Source : CS231n

En général, les coefficients (ou poids) d'un réseau de neurones sont actualisés avec une descente de gradient qui considère un groupe d'observations (*batch*). En effet, considérer un seul élément est lent et peut donner des gradients très bruités, mais à l'inverse, le jeu de données entier est en général trop volumineux pour être considéré en une seule fois, d'où la séparation en lots de données. Les différentes étapes sont alors les suivantes :

1. Prendre un groupe d'observations appartenant aux données du jeu d'entraînement.
2. Réaliser la propagation avant pour obtenir la fonction de coût correspondante.
3. Effectuer une rétropropagation de la fonction de coût pour obtenir les gradients.
4. Utiliser les gradients pour actualiser les coefficients du réseau.

Ces étapes sont répétées jusqu'à ce que tous les éléments aient été vus une fois par le réseau ; on appelle cela **une époque** (*epoch*). À la fin d'une époque, tous les éléments sont mélangés et une nouvelle époque est relancée.

Optimisation Tel que mentionné précédemment, l'actualisation des poids du modèle peut se faire par la méthode SGD. Toutefois, le problème de cette méthode

concerne les minimums locaux ou les points de selles, qui vont mener à des solutions sous optimales. Pour remédier à ce problème, d'autres méthodes utilisant un élan (*momentum*) ont été proposées et sont encore très étudiées. Parmi ces méthodes, on peut citer SGD + momentum, AdaGrad [31], RMSProp [56] ou encore Adam [72], très utilisé actuellement (voir figure 3.11).

Méthode	Explication	Mise à jour de w
Momentum	• Amortit les oscillations • Amélioration par rapport à la méthode SGD • 2 paramètres à régler	$w - \alpha v_{dw}$
RMSprop	• Root Mean Square propagation • Accélère l'algorithme d'apprentissage en contrôlant les oscillations	$w - \alpha \frac{dw}{\sqrt{s_{dw}}}$
Adam	• Adaptive Moment estimation • Méthode la plus populaire • 4 paramètres à régler	$w - \alpha \frac{v_{dw}}{\sqrt{s_{dw}} + \epsilon}$

FIGURE 3.11 – Différentes méthodes utilisées lors de l'optimisation du modèle. SGD : descente de gradient stochastique. $v_{dw} = \beta_1 v_{t-1} + (1 - \beta_1) dw$ et $s_{dw} = \beta_2 s_{t-1} + (1 - \beta_2) dw^2$ avec β_1 et β_2 les taux de décroissance des 1er et 2ème moments. Source : adaptée de CS230, avec l'autorisation de Afshine Amidi et Shervine Amidi, 2018

Les CNN ayant été définis, la section suivante expose des architectures classiques pour la segmentation d'images.

3.4 CNN pour la segmentation d'images médicales

Cette partie présente les architectures de base pour la segmentation d'images médicales ainsi que les fonctions de coût usuelles.

3.4.1 Architectures

Les réseaux de type encodeur-décodeur

Les encodeur-décodeurs sont utilisés pour des tâches de segmentation d'images, c'est-à-dire pour délimiter des régions d'intérêt ou attribuer une classe à chacun des pixels d'une image.

L'architecture d'un encodeur-décodeur est composée de deux parties distinctes :

— un encodeur : il extrait des attributs visuels et sémantiques de l'image d'entrée en compressant la représentation, jusqu'à un espace latent ;

— un décodeur : il reconstruit progressivement les cartes de caractéristiques jusqu'à la résolution d'entrée.

Un exemple d'encodeur-décodeur est présenté figure 3.12 : une image IRM 2D est donnée en entrée du réseau, qui va générer en sortie une carte de segmentation de l'image d'entrée par zones d'intérêt, ici les constituants du genou (os et cartilage).

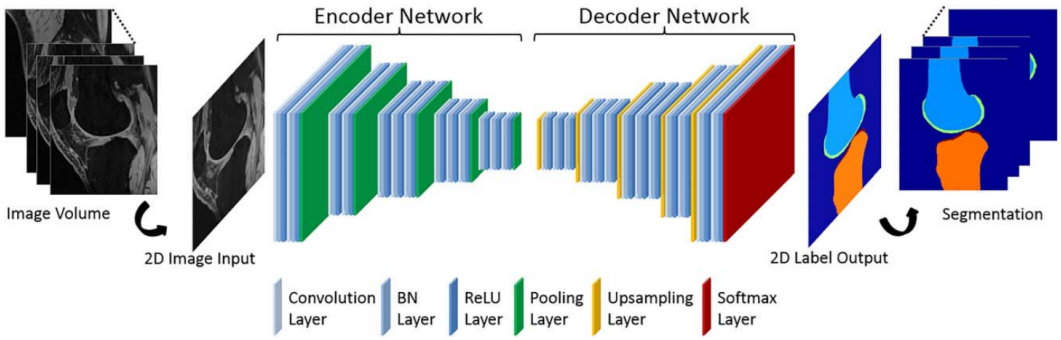


FIGURE 3.12 – Segmentation d'images IRM du genou grâce à un encodeur-décodeur. Source : Liu et al., 2018

U-Net : le standard en imagerie médicale Le réseau U-Net [102] a été développé pour répondre à des problèmes de segmentations d'images médicales. Fort de son succès (environ 30 000 citations en août 2021), il est également utilisé pour des applications dans d'autres domaines que le médical.

Le réseau tient son nom de sa forme : une partie qui encode l'information et une partie qui la décode jusqu'à revenir à la résolution originale de l'image, comme un encodeur-décodeur classique (voir figure 3.13). Sa spécificité réside dans la partie décodeuse : en plus du ré-échantillonnage des cartes de caractéristiques du décodeur à chaque bloc, celles de la partie encodeuse y sont concaténées grâce à des connexions résiduelles. De cette manière, des caractéristiques extraites à différentes résolutions peuvent être réutilisées lors de la reconstruction et donner de l'information sur le contexte spatial.

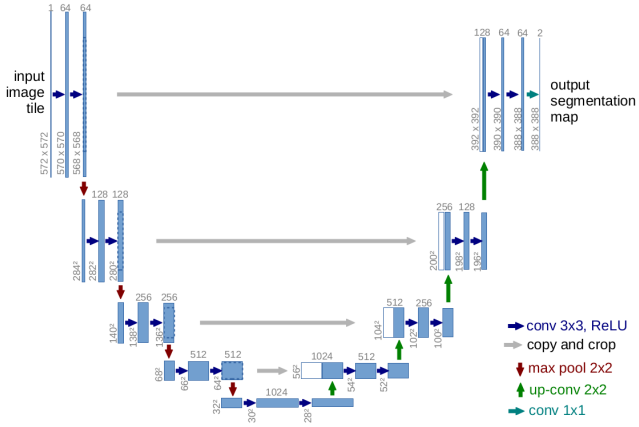


FIGURE 3.13 – Architecture du réseau U-Net. Source : Ronneberger et al. [102]

Les réseaux de type encodeur-décodeur correspondent actuellement au standard en segmentation d'images médicales. Ils servent également de base à d'autres architectures plus complexes, dont il existe de nombreuses variantes (réseaux en cascade, multitâche, avec une branche adversaire, etc.).

3.4.2 Fonctions de coût

En apprentissage profond, l'équation 3.5 définissant la fonction de coût à optimiser prend la forme suivante :

$$R(W) = \frac{1}{N}\sum_i R_i\big(f_W(\mathbf{x_i}), y_i\big) + \lambda R_{struct}(W) \tag{3.12}$$

avec $R_{struct}(W)$ le terme de régularisation ne dépendant que des poids du réseau. Plusieurs fonctions peuvent être utilisées comme régularisation, parmi lesquelles la régularisation L1 (*lasso*) ou L2 (*Ridge*) ou encore la somme des deux :

— Régularisation L1 : $R(W) = \sum_k \sum_l |W_{k,l}|$

— Régularisation L2 : $R(W) = \sum_k \sum_l W_{k,l}^2$

En pratique, la régularisation L2 est la plus utilisée.

Dans cette partie, nous détaillons les fonctions de coût les plus courantes pour les modèles de segmentation. Il existe toutefois de nombreuses variantes adaptées aux spécificités des tâches de segmentation.

3.4.2.1 Entropie croisée

L'entropie croisée est couramment utilisée comme fonction de coût pour entraîner un réseau de neurones à une tâche de segmentation ou de classification.

L'entropie croisée $\mathcal{H}$ entre une distribution cible p et une distribution estimée q est définie par :

$$\mathcal{H}(p,q) = -\sum_x p(x)\log q(x). \tag{3.13}$$

L'entropie croisée peut également être exprimée en fonction de la divergence de Kullback-Leibler (notée D_{KL}), mesure de dissimilarité entre deux distributions de probabilités :

$$\mathcal{H}(p,q) = \mathcal{H}(p) + D_{KL}(p||q). \tag{3.14}$$

Minimiser l'entropie croisée revient donc à minimiser la divergence de Kullback-Leibler.

En considérant les probabilités en sortie de la fonction softmax (voir section 3.3.1.5), l'entropie croisée utilisée pour l'entraînement d'un réseau correspond à la formule suivante :

$$EC(x,y) = -\sum_i^N \sum_c^C y_{ci}\log p_{ci} \tag{3.15}$$

où $p_{ci} = p(Y = c|\mathbf{x_i})$ est la probabilité prédite par le modèle pour l'observation i d'appartenir à la classe c et $y_{ci} = 1$ si $y_i = c$ et 0 sinon (la vérité terrain encodée sous forme *one-hot* pour le pixel i).

Entropie croisée partielle La formule de l'entropie croisée ci-dessus suppose que la vérité terrain soit entièrement connue (l'étiquette de l'image pour un problème de classification ou la classe de chacun des pixels pour un problème de segmentation). Elle peut être adaptée à des problèmes faiblement supervisés, où une partie seulement de la vérité terrain est connue en ne considérant que les pixels annotés, comme proposé dans Tang et al. [123] :

$$EC(x,y) = -\sum_{i\in\Omega_L} \sum_c^C y_{ci}\log p_{ci} \tag{3.16}$$

avec Ω_L la segmentation partielle de l'image.

3.4.2.2 La fonction de coût Dice généralisée

La métrique de similarité de Dice est classique pour les problèmes de segmentation (voir l'équation 3.23). Plusieurs fonctions de coût dérivables ont été proposées à partir de cette métrique, dont la fonction de coût Dice généralisée [116]. Dans un problème multiclasse à C classes avec une pondération sur les classes, l'expression est la suivante :

$$\text{Dice loss} = 1 - 2\frac{\sum_c^C w_c \sum_{i=1}^N y_{ci} p_{ci} + \epsilon}{\sum_c^C w_c \sum_{i=1}^N (y_{ci} + p_{ci}) + \epsilon} \tag{3.17}$$

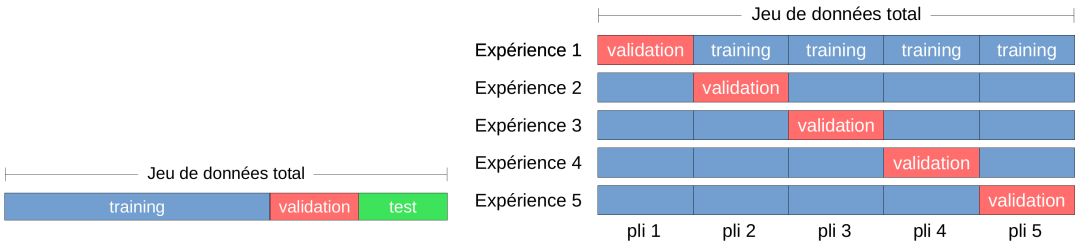
où w_c est la pondération sur la classe c, p_{ci} la probabilité prédite par le modèle pour l'observation i d'appartenir à la classe c et y_{ci} est égal à 1 si le pixel i appartient à la classe c, 0 sinon (la vérité terrain encodée sous forme *one-hot* pour le pixel i).

3.5 Évaluation des performances

Dans cette partie, il s'agira de présenter la mise en place d'expériences et le calcul de métriques pour une évaluation correcte des performances d'un modèle.

3.5.1 Stratégie d'évaluation

Pour mener à bien des expériences, il s'agit tout d'abord d'organiser correctement les données. En effet, comme évoqué à la section 3.2.1, une partie des données à disposition doit servir à entraîner le modèle et une autre à l'évaluer. Le test sur de nouvelles données est indispensable, pour s'assurer que le modèle généralise correctement et n'a pas sur-appris (voir figure 3.2).



(A) Train/Valid/Test (B) Validation croisée à 5 plis (ou *folds*).

FIGURE 3.14 – Différents types d'entraînement possibles. (A) Un seul modèle est entraîné sur les données d'entraînement (train) et sélectionné grâce au jeu de validation (val), plus petit. Il est ensuite testé sur un jeu indépendant de test. (B) En validation croisée à 5 plis, chaque pli - de taille égale - est tour à tour utilisé comme ensemble de validation. 5 modèles sont alors obtenus dont les performances sont moyennées.

Train / Validation / Test Généralement, le jeu de données est séparé en deux ensembles : un ensemble pour l'entraînement (*train*), et un pour le test. L'ensemble d'entraînement est lui-même séparé en 2, avec un ensemble de validation dont les exemples ne sont pas utilisés pour apprendre les poids du modèle mais sur lesquels on évalue les performances, pour choisir au mieux les hyperparamètres. L'évaluation finale se fait sur le jeu de test, et ne doit être faite qu'une fois.

Validation croisée Dans le cas où la quantité de données est trop faible pour avoir suffisamment d'exemples dans l'ensemble de validation (ce qui est souvent le cas en imagerie médicale), la stratégie de validation croisée à k plis est à privilégier (figure 3.14B). L'idée est qu'au lieu de prendre les n premiers exemples pour le jeu de validation et le reste pour le jeu d'entraînement, on peut obtenir une estimation meilleure et moins bruitée des performances du modèle avec les hyperparamètres choisis en itérant sur différents ensembles de validation, et en moyennant ensuite les performances obtenues. Par exemple, pour une validation croisée à 5 plis, les 4 premiers plis seront utilisés pour l'entraînement et le 5ème servira à la validation. Chaque pli sera utilisé tour à tour comme pli de validation et les performances finales calculées en faisant la moyenne des performances obtenues sur chacun des plis.

En plus de la validation croisée, une partie des données peut être gardée de côté pour constituer un jeu de test inutilisé lors de l'optimisation du modèle.

3.5.2 Métriques

Les métriques utilisées dépendent de la tâche évaluée et sont adaptées à cette-dernière : classification binaire, classification multiclasse, détection ou encore segmentation.

3.5.2.1 Classification binaire

Dans une tâche de classification binaire, il s'agit d'assigner une classe (0 ou 1) à chacun des exemples donnés. Quatre cas peuvent se présenter, illustrés figure 3.15 :

— une classe négative prédite négative (VN)

— une classe négative prédite positive (FP)

— une classe positive prédite positive (VP)

— une classe positive prédite négative (FN)

Les principales métriques calculées à partir de ces quantités sont formulées figure 3.16. Une présentation sous la forme d'une matrice de confusion permet d'avoir une image complète des performances du modèle (voir figure 3.15).

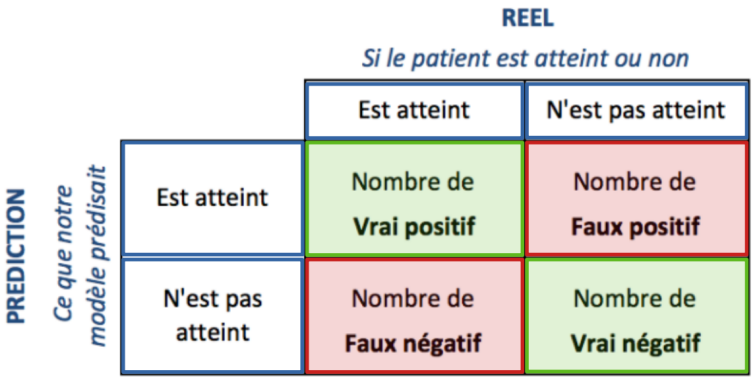


FIGURE 3.15 – Matrice de confusion pour un problème de classification binaire.

Indicateur	Formule	Interprétation
Accuracy	$\frac{VP+VN}{VP+VN+FP+FN}$	Performances globales du modèle
Précision	$\frac{VP}{VP+FP}$	À quel point les prédictions positives sont précises
Rappel ou sensibilité	$\frac{VP}{VP+FN}$	Couverture des observations vraiment positives
Spécificité	$\frac{VN}{VN+FP}$	Couverture des observations vraiment négatives
F-mesure	$\frac{2VP}{2VP+FP+FN}$	Indicateur hybride utilisé pour les classes déséquilibrées

FIGURE 3.16 – Principales métriques pour évaluer les performances des modèles de classification. VP : Vrai Positif, VN : Vrai Négatif, FP : Faux Positif, FN : Faux Négatif. Inspirée de CS229.

Les courbes ROC Les courbes ROC (pour *Receiver Operating Characteristic*) sont très utilisées pour évaluer des classifieurs. On y représente le taux de vrais positifs (TVP, équivalent au rappel ou à la sensibilité) en fonction du taux de faux positifs (TFP) - tous deux bornés à 1 - à différents seuils de décision possibles (figure 3.17).

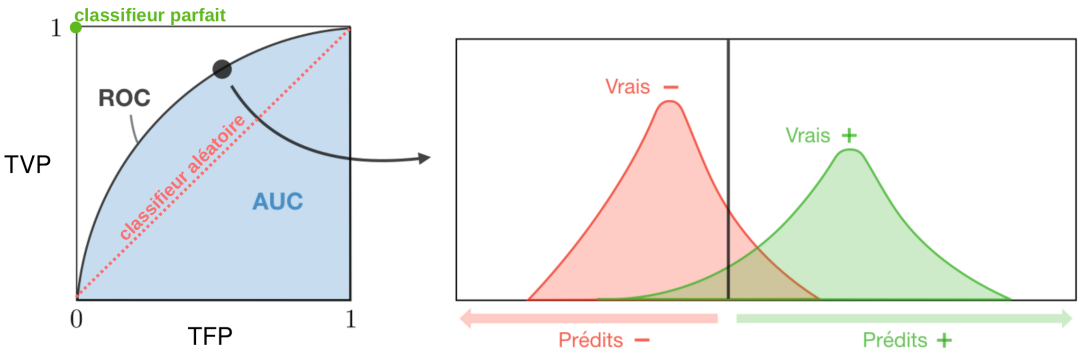


FIGURE 3.17 – Courbe ROC et aire sous la courbe (AUC). Un classifieur aléatoire correspondrait à la droite $y = x$ et un classifieur parfait aurait un TVP à 1 avec un TFP à 0. Plus l'AUC est grande, meilleur est le modèle. Adaptée de CS229, avec l'autorisation de Afshine Amidi et Shervine Amidi, 2018

Leur formule est rappelée ici :

$$\text{TVP} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{3.18}$$

$$\text{TFP} = \frac{\text{FP}}{\text{VN} + \text{FP}} = 1 - \text{spécificité}. \tag{3.19}$$

L'intérêt de la courbe ROC dans le domaine médical a été souligné depuis une quarantaine d'années [91, 121]. La courbe ROC permet la comparaison des performances diagnostiques de plusieurs tests à l'aide de l'évaluation des aires sous la courbe (AUC, pour *Area Under the Curve*). Elle est aussi utilisée pour estimer la valeur seuil optimale d'un test en tenant compte de critères spécifiques au besoin clinique. Il faut noter que le tracé des courbes ROC s'applique à la détection d'une seule anormalité dans l'image sans aucune indication sur sa localisation.

3.5.2.2 Classification multiclasse

Matrice de confusion La matrice de confusion présentée dans le cas binaire (voir figure 3.15) peut facilement être adaptée au cas multiclasse, toujours en mettant en regard la classe prédite à la vraie classe (voir figure 3.18). Dans ce cas, les VP correspondent aux termes dans la diagonale.

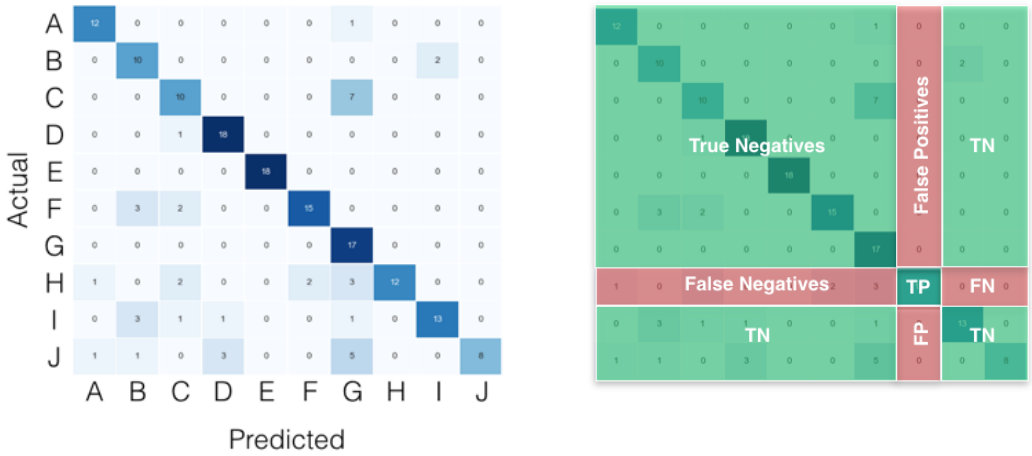


FIGURE 3.18 – Matrice de confusion et interprétation dans le cas de classification multiclasse. Un classifieur parfait a des valeurs non nulles dans la diagonale seulement. Source : it swarm, licence CC BY-SA 3.0.

Score kappa de Cohen Le score kappa de Cohen (du nom de son inventeur) est une métrique quantitative pouvant être extraite des matrices de confusion. Il mesure l'accord entre deux observateurs en éliminant la part de hasard dans l'accord. Ce score est un nombre réel compris entre -1 et 1, l'accord maximal correspondant à un kappa de 1. Sa formule est donnée équation 3.20 :

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{3.20}$$

avec P_0 la proportion de l'accord observé entre les deux observateurs et P_e la probabilité d'un accord aléatoire.

Accord	Kappa
Presque parfait	0.81 - 1
Bon	0.61 - 0.80
Modéré	0.41 - 0.60
Faible	0.21 - 0.40
Très faible	0.0 - 0.20
Désaccord	<0

TABLEAU 3.1 – Degré d'accord selon la valeur de kappa proposé par Landis et al. [73].

Dans le cas où les classes sont ordonnées, c'est-à-dire qu'il existe une hiérarchie entre elles, le score de kappa peut également être pondéré de manière à pénaliser davantage des erreurs plus importantes. Dans ce cas, une matrice de pondération

est définie pour pondérer chacun des termes de la matrice de confusion. Cette pondération peut être linéaire, où chaque poids de la matrice se calcule d'après :

$$w_{ij} = 1 - \frac{|i-j|}{r-1} \tag{3.21}$$

ou bien quadratique :

$$w_{ij} = 1 - \frac{(i-j)^2}{(r-1)^2} \tag{3.22}$$

avec i la i^{ème} colonne de la matrice de poids, j la j^{ième} ligne et r le nombre de modalités de jugement.

3.5.2.3 Détection

Les courbes FROC Les courbes FROC [42, 12] sont une variante des courbes ROC (d'où leur nom *Free-response* ROC), appropriées à des tâches de détection.

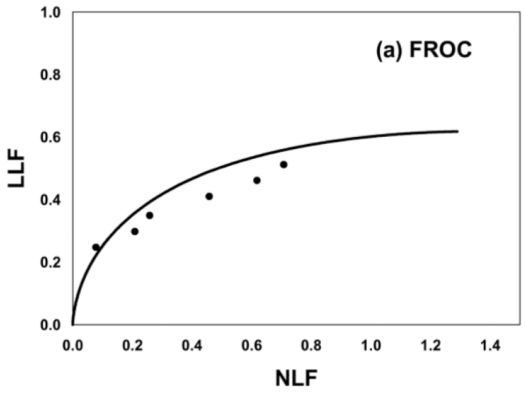


FIGURE 3.19 – Courbe FROC, où l'axe des abscisses est non borné contrairement aux courbes ROC. LLF : *Lesion Localization Fraction* ou sensibilité, NLF : *Non-lesion Localisation Fraction* ou nombre de faux positifs (souvent par patient ou par image). Source : Chakraborty [18].

Il s'agit - pour l'observateur ou le modèle - de localiser et d'évaluer toutes les anomalies présentes dans l'image sans limites de nombre, en indiquant un niveau de confiance pour chacune. L'axe des abscisses correspond au nombre moyen de faux positifs (FP) par patients ou images, au lieu du taux de faux positifs (FP/N, où N est le nombre de négatifs) pour les ROC. La spécificité des courbes FROC est donc de considérer le nombre d'anormalités repérées et leur localisation en plus de la précision du modèle. Les courbes FROC permettent d'extraire une information plus proche de la réalité clinique et plus pertinente des performances d'un modèle de détection, localisation et classification d'une ou plusieurs anormalités chez un sujet.

3.5.2.4 Segmentation

L'indice de Sørensen-Dice L'indice Sørensen–Dice (aussi appelé F-score) est un indicateur pour mesurer la similarité de deux ensembles. Il est très largement utilisé en imagerie médicale comme mesure de qualité d'une segmentation. Pour deux ensembles X et Y, le Dice est donné par la formule suivante :

$$Dice(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}. \tag{3.23}$$

Une segmentation parfaite, où les deux ensembles se recouvrent entièrement, correspond à un Dice de 1. À l'inverse, une segmentation sans recouvrement correspond à un Dice de 0 (exemples figure 3.20)

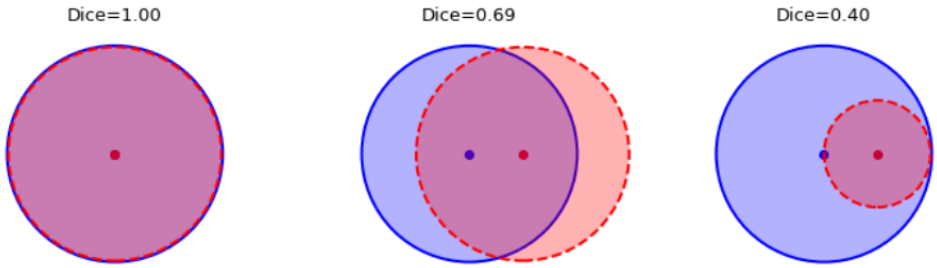


FIGURE 3.20 – Valeurs de Dice pour plusieurs exemples. Source : il-monteux.github.io

3.6 Conclusion

Dans ce chapitre, nous avons abordé les fondements de l'apprentissage profond. Nous avons présenté les bases de l'apprentissage automatique, les architectures classiques des réseaux de segmentation, les fonctions de coûts les plus utilisées en segmentation d'images et la manière d'évaluer un modèle.

Dans le chapitre suivant, nous présentons un état de l'art des systèmes d'aide au diagnostic (CAD) pour la détection ou caractérisation du cancer de la prostate utilisant des techniques d'apprentissage profond.

Chapitre 4

Les systèmes d'aide au diagnostic (CAD) pour l'imagerie du cancer de la prostate

Sommaire

4.1 Introduction

Depuis une cinquantaine d'années, les systèmes d'aide au diagnostic (CAD) sont très étudiés, pour des applications médicales variées. Un CAD est un programme informatique capable d'aider le radiologue dans l'interprétation d'images médicales. Cela peut consister à identifier des anomalies (tumeurs par exemple), à prédire leur évolution ou encore à classer les images en différentes catégories. On peut distinguer les systèmes d'aide à la décision (CADx), qui vont classer une région d'intérêt préalablement identifiée, des systèmes d'aide à la détection (CADe), qui vont localiser les anomalies dans l'image.

En 1998, le premier système CAD commercial pour la détection de microcalcifications en mammographie, ImageChecker de R2 technologies, a été approuvé par la *Food and Drug Administration* américaine (FDA). Dans les années qui ont suivi, plusieurs CAD ont aussi été approuvés par la FDA pour diverses applications : analyse d'images de poumons, du colon et du cœur.
L'avènement de l'intelligence artificielle et plus spécifiquement de l'apprentissage profond a permis d'améliorer les performances des systèmes CAD, jusqu'à égaler les performances des experts pour certaines tâches.

Fin 2016, des ingénieurs de Google publiaient une étude sur le test d'un algorithme d'apprentissage profond pour la détection des rétinopathies diabétiques [55]. Entraîné sur une base de plus de 128 000 images de fonds d'œil, le réseau a produit

un diagnostic correct avec des résultats comparables à ceux obtenus par des ophtalmologistes bien entraînés, sur 2 jeux de données différents.

En janvier 2017, une équipe de Stanford faisait la une de la revue Nature en obtenant d'excellents résultats avec un GoogleNet Inception v3 pour la distinction de grains de beauté et mélanomes : 90 % des taches bénignes sur la peau ont été correctement classifiées par le réseau, contre 76 % pour les dermatologues interrogés sur environ 130 000 images de lésions cutanées [45].

Toutefois, les CAD ne sont pas développés dans l'optique de remplacer les experts, mais bien de les assister pour améliorer leurs performances. Les systèmes sont développés pour une tâche bien précise et n'ont pas les capacités de raisonnement, d'analyse et de communication propres au cerveau humain.

En outre, tous les développements de CAD ne sont pas des succès et l'application de l'intelligence artificielle pour la santé est rarement une tâche simple.

Récemment, en 2021, l'entreprise IBM a mis en vente sa filiale d'intelligence artificielle médicale Watson Health créée en 2015, après avoir beaucoup misé dessus. Pourtant, des études publiées par le groupe montraient que les performances de leur système *Watson for Oncology*, IA spécialisée dans le traitement des cancers, approchaient celles des professionnels.

Les systèmes CAD sont toujours en cours d'étude et de développement pour de nombreuses applications. Par la suite, nous nous concentrerons sur les systèmes CAD appliqués à l'IRM de prostate et utilisant des méthodes d'apprentissage profond ([131, 126]).

4.2 Les CAD pour la segmentation de la prostate en IRM

La segmentation de la prostate sur des images IRM est encore une tâche très étudiée. Un état de l'art condensé des papiers publiés depuis 2019 est présenté tableau 4.1. Un état de l'art détaillé est disponible dans des papiers de revue récents [51, 71].

De manière générale, la segmentation de la prostate se fait à partir de la séquence en T2-w axial, où l'anatomie est la plus visible (voir section 2.7.2). Les réseaux utilisés varient, mais la plupart sont inspirés du U-Net 2D. Les fonctions de coûts les plus utilisées sont la Dice loss et l'entropie croisée, ou bien la somme des deux.

La comparaison entre les différentes méthodes n'est pas immédiate, compte tenu des différences en termes de base de données (nombre de patients, vérité terrain, homogénéité de la base en termes de scanners, sites, etc.). D'ailleurs, un certain nombre de papiers sont évalués sur des bases de données privées qui ne sont donc pas disponibles.

Toutefois, nous pouvons extraire des tendances de l'analyse du tableau 4.1. La tâche de segmentation de la prostate obtient un Dice supérieur à 90 % dans la majorité des études publiées. L'étude de Jin et al. [68] rapporte les meilleurs Dices avec une valeur supérieure à 96 % lorsque leur 3D PBV-Net est testé sur la base de données publique PROMISE12 [83]. Pour la segmentation de la ZP et de la ZT (ou de la glande centrale, notée GC), les performances sont un peu en deçà : les valeurs de Dice rapportées sur la ZP varient entre 71 et 92 % et entre 87 et 94 % pour la ZT. La segmentation de la ZP semble donc être la tâche la plus ardue.

De nos jours, l'une des principales limites pour l'utilisation de CAD en clinique concerne leur capacité de généralisation. En effet, un modèle dont les performances

sont excellentes sur un certain jeu de données peut les voir chuter considérablement lorsqu'il sera testé sur un jeu de données différent. C'est notamment pour cette raison qu'il est important d'entraîner les modèles sur des jeux de données hétérogènes afin qu'ils soient plus robustes à de nouveaux jeux de données. Cette hétérogénéité peut être présente par l'inclusion de scanner de marques différentes, avec des champs ou paramètres d'acquisition variables, ou bien par l'acquisition des données sur des sites différents, avec des protocoles potentiellement distincts. Le tableau comparatif établi ci-dessous rend compte de l'émergence de cette problématique. Zavala-Romero et al. [140] entraînent leur modèle selon différentes configurations : en incluant seulement le scanner Siemens 3T, en incluant seulement le scanner GE Discovery 3T ou bien en incluant les deux simultanément. Leurs expériences montrent que les performances du réseau entraîné sur l'un des deux scanners sont nettement inférieures lorsqu'il est testé sur un nouveau scanner. Les meilleurs résultats sont obtenus en mélangeant les deux types de scanners dans les données d'entraînement. Par ailleurs, Rundo et al. [108] et Liu et al. [85] mettent l'accent dès le titre de leur article sur le caractère hétérogène des bases de données inclues dans leur étude, pour en valider la robustesse.

En conclusion, la segmentation de la prostate et de ses différentes zones anatomiques est encore un sujet d'actualité, mais des performances très correctes commencent à être rapportées. Toutefois, la robustesse de ces modèles sur des données externes et leur capacité de généralisation n'est pas toujours abordée et doit être mise au centre des futures études. En outre, une base de données publique hétérogène et de grande envergure est toujours manquante pour une évaluation juste entre les différents systèmes.

4.3 Les CAD pour la détection, la segmentation et la caractérisation du cancer de la prostate en IRM

La détection du CaP sur des IRM est une tâche d'un grand intérêt clinique. En conséquence, de nombreuses études s'attelant à ce problème ont été publiées ces dernières années. Plusieurs papiers de revue en font l'inventaire, les plus récents et complets étant Wildeboer et al. [131] et Twilt et al. [126].

4.3.1 Approches par apprentissage classique et historique à CREATIS

Bien que de plus en plus de travaux utilisent des techniques d'apprentissage profond, les CAD pour le CaP basés sur des techniques d'apprentissage automatique classiques sont nombreux et étudiés depuis plusieurs dizaines d'années [78, 84, 46, 17]. Dans ces approches, des caractéristiques sont extraites manuellement avant d'entraîner un classifieur.

A CREATIS, cette problématique est également étudiée depuis une quinzaine d'années. Dans Niaf et al. [96], 4 méthodes d'apprentissage sont comparées : les séparateurs à vaste marge (noté SVM, de *Single Vector Machine*), l'analyse linéaire discriminante, les k plus proches voisins et les classifieurs naïfs de Bayes. Sur une sous-partie de la base utilisée dans cette thèse (30 patients), l'approche par SVM s'avère la plus performante avec une AUC de 0.89 [0.81–0.94] pour la distinction des tissus malins et bénins issus de la ZP à partir de 15 caractéristiques issues des niveaux de gris des IRM-mp telles que les caractéristiques de Haralick, de gradient, etc. Une étude ultérieure [95] a montré que l'utilisation de ce système CAD par

les radiologues leur permettait d'améliorer leur performance, sans différence significative toutefois. Par la suite, des approches par apprentissage de dictionnaires ont été étudiées [77]. Enfin, les derniers travaux au sein du laboratoire sur le CaP se sont penchés sur la question de l'adaptation de domaine entre les différentes sources constituées par les différents scanners [5, 49].

Par la suite, nous tenterons de faire un état de l'art le plus exhaustif possible des travaux concernant la détection ou segmentation des lésions cancéreuses dans l'IRM de prostate avec des méthodes d'apprentissage profond supervisé. Les travaux concernant la classification binaire de régions d'intérêts (systèmes CADx) ne seront pas abordés.

4.3.2 CADe pour la détection ou segmentation binaire du CaP

Dans cette partie, on considère les systèmes CADe binaires, c'est-à-dire des modèles qui vont localiser voire segmenter les lésions dans l'image donnée en entrée, contenant toute la prostate.

Le tableau 4.2 présente les algorithmes pour la détection ou segmentation binaire des lésions publiés dans des journaux entre janvier 2018 et juillet 2021.

Encore une fois, la comparaison n'est pas immédiate compte tenu des différences en termes de bases de données (souvent privées) mais également dans les métriques et critères d'évaluation. Globalement, en considérant les 9 papiers listés :

- Les bases comprennent une centaine de patients, souvent acquis sur un seul scanner ou plusieurs modèles d'un même constructeur (6/8 papiers, 1 indéterminé).
- La modalité T2-w est toujours utilisée en entrée, seule quelquefois [61, 6] mais majoritairement avec la diffusion, sous forme de carte paramétrique ADC [130, 26] ou avec l'ADC et la diffusion acquise à une haute valeur de b (voir la définition section 2.7.3) [117, 135, 111, 22, 109]. Aucune étude basée sur des méthodes d'apprentissage profond n'utilise l'imagerie de perfusion à notre connaissance ; les données utilisées sont donc biparamétriques.
- La majorité des études ont une vérité terrain qui n'est pas parfaite, car basée sur le PI-RADS ou la biopsie, la pièce de prostatectomie n'étant que très rarement disponible (1/9 papier).
- Les lésions incluses sont souvent celles jugées comme cliniquement significatives, majoritairement définies comme de GG ≥ 2 [130, 111, 109].
- L'entraînement et la validation des modèles se fait soit par séparation du jeu en ensembles d'entraînement, de validation et de test (7/9 papiers), soit par validation croisée à 4 ou 5 plis (2/9 papiers).
- Les métriques rapportées sont la sensibilité, l'AUC, le Dice et une fois seulement le score κ de Cohen (voir la définition des métriques section 3.5.2).
- La majorité des travaux [61, 6, 135, 111] utilisent des réseaux classiques (de type U-Net, Res-Net) avec les fonctions de coût les plus courantes en segmentation (Dice loss et entropie croisée, souvent pondérées compte tenu du déséquilibre de classes).

Le travail présenté dans Schelb et al. [111], où un U-Net est entraîné sur une base de 312 patients, est très encourageant puisqu'il montre que les performances du réseau sont similaires au critère PI-RADS utilisé par les radiologues (voir section 2.5.2) pour identifier les lésions CS. En considérant les sextants (zones biopsiées) obtenant

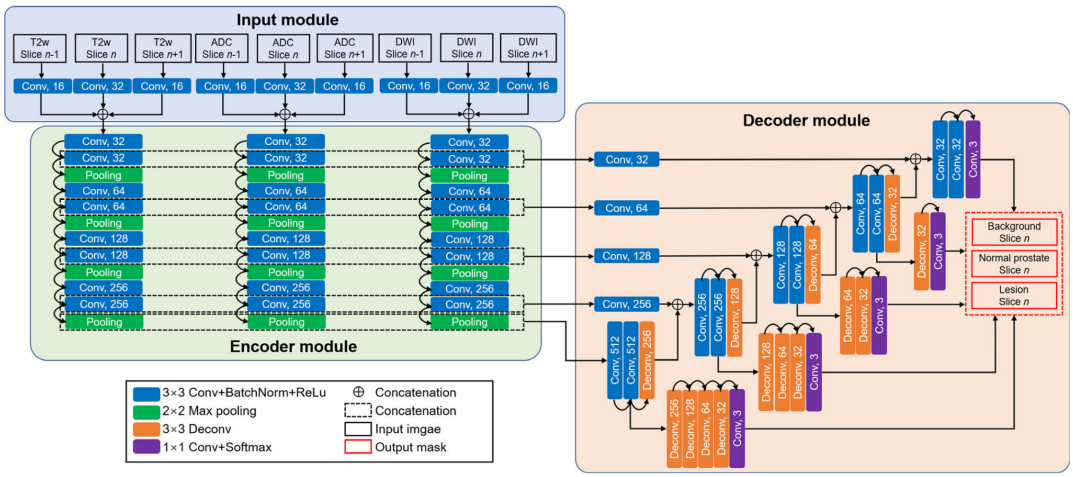


FIGURE 4.1 – Architecture du modèle multibranches proposé par Chen et al. [22].

une probabilité de malignité ≥ 0.22 et ≥ 0.33 avec U-Net, les performances ne sont pas significativement différentes des critères PI-RADS ≥ 3 et PI-RADS ≥ 4 respectivement.

Certains modèles sont toutefois plus originaux. Wang et al. [130], proposent un modèle basé sur des GoogleNet pré-entraînés sur ImageNet pour la détection binaire des lésions (voir figure 8.9). La spécificité de l'algorithme est l'entraînement joint de deux réseaux, prenant chacun une modalité (T2-w ou ADC) en entrée. Une fonction de coût de cohérence entre les modalités est ajoutée à l'entropie croisée associée à chacune des branches. Cette idée a d'ailleurs été reprise par Cao et al. [15] (tableau 4.4) pour la détection multiclasse des lésions par GS : une fonction de coût "d'observation mutuelle" (*mutual finding*) est utilisée en plus de la fonction de coût focale pour entraîner un CNN à 2 branches (une par modalité).

Chen et al. [22] ont proposé un réseau basé sur le U-Net mais avec plusieurs branches (voir figure 4.1) : l'encodeur du réseau est composé de 3 branches, avec une branche par modalité (T2-w, ADC et DWI) pour une fusion tardive. Pour chaque branche, les coupes précédant et suivant la coupe d'intérêt sont données en entrée, donnant une approche en 2.5D. Ce type d'approche en 2.5D a déjà été utilisé dans d'autres applications médicales et appliqué au CaP dans Alkadi et al. [6]. En plus des différentes branches dans l'encodeur, le réseau a 4 sorties à différents niveaux du décodeur : 2 sorties dans des couches profondes effectuent la segmentation des lésions, une sortie à un niveau intermédiaire segmente la prostate et une dernière sortie à la résolution initiale détermine l'arrière-plan. Les sorties sont ré-échantillonnées à la résolution initiale (sauf la dernière) et utilisées dans le calcul de l'entropie croisée.

Enfin, très récemment Saha et al. [109] ont proposé un modèle composé de deux CNN entraînés en parallèle : un U-Net 3D pour la segmentation des lésions joint à un classifieur d'imagette 3D, permettant de réduire les faux positifs (voir figure 4.2). Une originalité de leur travail réside également dans l'inclusion d'un *a priori* anatomique probabiliste concernant la localisation de la prostate et des lésions, calculé à partir de leur base de données. Cette carte de probabilités est utilisée en fusion précoce comme un canal supplémentaire en entrée du réseau. Leur modèle est entraîné sur une base conséquente (2436 patients) mais à la vérité terrain peu fiable, car basée sur le critère PI-RADS donné par les radiologues, qui ne permet l'inclusion que des lésions les plus visibles à l'IRM. Les lésions de score PI-RADS ≥ 4 sont considérées comme cliniquement significatives. Toutefois, le modèle est validé sur une

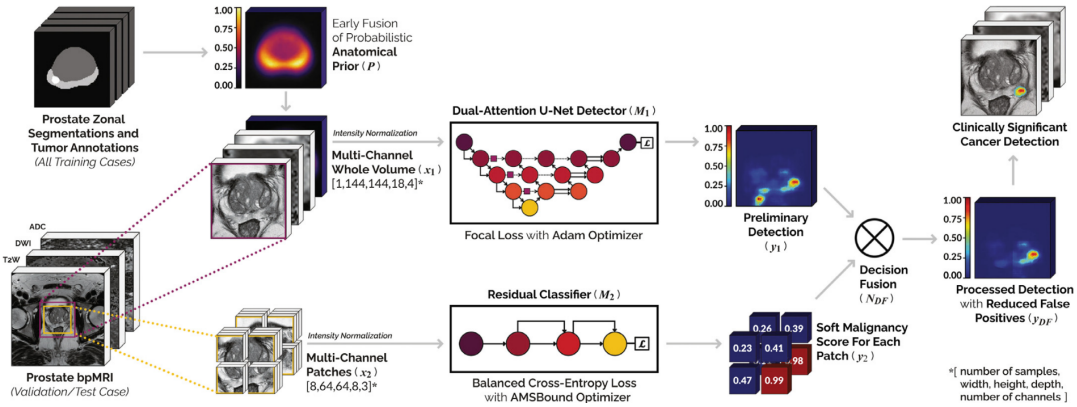


FIGURE 4.2 – Modèle proposé par Saha et al. [109].

base externe plus fiable composée de 296 patients, dont la vérité terrain est basée sur la biopsie. La notion de lésion cliniquement significative est donc plus sûre mais les lésions biopsiées restent celles qui ont été repérées à l'IRM. Les performances rapportées sont élevées, avec environ 84 % de sensibilité à 0.5 FP par patient sur la base d'entraînement et environ 77 % de sensibilité sur le jeu externe. Cette baisse de performance sur le jeu externe est potentiellement due à la vérité terrain plus fiable (biopsie au lieu du PI-RADS) et / ou à un problème d'adaptation de domaine.

Cette idée exploitée par Saha et al. [109] qu'un contexte anatomique pourrait aider le réseau à détecter les lésions a également été utilisée dans d'autres travaux, où la prostate est segmentée en plus des lésions [6, 111, 26, 22].

4.3.3 CADx pour la classification multiclasse du CaP

Les systèmes CADx pour la classification multiclasse sont des modèles qui, à partir d'une région d'intérêt centrée sur une lésion, vont prédire un score de Gleason (GS) associé.

En 2017, le challenge ProstateX-2 [82] a été proposé pour classer des régions d'intérêt (lésions) par GS. Les gagnants [3] ont proposé une méthode où des caractéristiques de haut niveau sont extraites à partir de caractéristiques de texture grâce à des auto-encodeurs puis classées à l'aide d'un *softmax*. Ils ont obtenu un score de kappa à pondération quadratique (voir la définition section 3.5.2.2) de 0.2326 en validation croisée à 3 plis sur les données d'entraînement du challenge et de 0.2772 lors de l'évaluation sur les 70 lésions du jeu de test.

Dans un papier plus récent [2], la même équipe a utilisé un VGG-16 suivi d'un classifieur ordinal (voir illustration de l'encodage ordinal figure 4.3) pour caractériser le CaP. Ils ont obtenu un score de kappa à pondération quadratique de 0.4727 dans une stratégie de validation croisée *Leave-One-Patient-Out* (LOPO) sur le jeu d'entraînement de ProstateX-2, correspondant au meilleur score rapporté sur les données du challenge. Les performances sur les données de test du challenge n'ont pas pu être calculées, le challenge étant clos et le jeu de test non publié.

Plus récemment, Chaddad et al. [16] ont proposé un nouveau modèle qui utilise des caractéristiques d'entropie obtenues grâce à des CNNs pour prédire le GS de lésions de prostate. Les caractéristiques issues des différents CNNs sont combinées et utilisées en entrée d'une forêt aléatoire qui discerne les lésions GS 6, 3+4, 4+3, 8 et $\geq$ 9 avec une AUC de 80.08, 85.77, 97.30, 98.20 et 86.51 respectivement, après entraînement et test sur le jeu d'entraînement de ProstateX-2.

4.3.4 CADe pour la détection ou segmentation multiclasse du CaP

Les systèmes CADe multiclasse considérés ici vont localiser les lésions dans une coupe donnée en entrée et également prédire un score d'agressivité pour chaque lésion.

À notre connaissance, seulement 3 études proposent à ce jour des modèles pour la détection ou segmentation des lésions par agressivité, c'est-à-dire par GS ou bien par PI-RADS. Ces études sont très récentes et ont été publiées après le démarrage de cette thèse. Elles sont détaillées tableau 4.4.

Cao et al. [15] ont été les premiers à proposer un réseau de neurones pour la détection et la prédiction du GS associé au CaP. Ils utilisent un réseau qu'ils nomment FocalNet (voir figure 4.3) : il s'agit d'un CNN résiduel multiclasse à 101 couches entraîné avec une fonction de coût focale - qui accorde plus d'importance aux exemples mal classés - en plus de la fonction de coût de "résultat mutuel" évoquée ci-dessus (section 4.3.2). Les données sont encodées de manière ordinale, c'est-à-dire en prenant en compte la hiérarchie entre les classes : une lésion GS 4+3 sera considérée comme appartenant également à la classe précédente GS 3+3. Le réseau produit une prédiction sur 5 canaux de l'agressivité du cancer. À partir de ces cartes à l'échelle du pixel sont extraits des points de localisation des lésions, correspondant donc à une tâche de détection et non de segmentation. Le modèle est entraîné par validation croisée avec une base privée de 417 patients imagés sur 4 modèles de Siemens 3T avec la prostatectomie comme vérité terrain. Basé sur les points de détection, FocalNet atteint environ 88% de sensibilité à 1 FP par patient en ne considérant que les coupes ayant au moins une lésion. Concernant la prédiction du GS, les AUC révèlent une bonne distinction entre GS$\geq$7 vs. GS<7 et GS$\geq$4+3 vs. GS$\leq$3+4 (0.81 ± 0.001 et 0.79 ± 0.001) mais sont plus faibles pour différencier les GS$\geq$8 vs GS<8 et GS$\geq$9 vs GS<9 (0.67 ± 0.004 et 0.57 ± 0.002), parmi les lésions détectées. Ce travail reste la référence pour la tâche de détection des lésions par agressivité. Toutefois, il s'agit ici de détection et non de segmentation des lésions et la généralisation du modèle n'est pas abordée.

De Vente et al. [27] ont proposé récemment un modèle pour la segmentation des lésions par GS, entraîné et validé sur la base de données publique ProstateX-2. Ils ont préalablement adapté la base de classification à une tâche de segmentation en délimitant manuellement les lésions dans la vérité terrain à partir des centroïdes fournis par le challenge. Le réseau utilisé est un U-Net 2D entraîné à une tâche de régression ordinale. Dans le papier, l'incorporation d'information concernant la zone de la prostate où se trouve la lésion a été étudiée mais n'a pas montré de bénéfice. Ils rapportent un score de kappa à l'échelle de la lésion de 0.172 ± 0.169 sur le jeu d'entraînement de ProstateX-2 avec une validation croisée à 5 plis et de 0.13 ± 0.27 sur l'ensemble de test.

4.3.5 Les limites actuelles des systèmes CAD

Les systèmes CAD appliqués à la prostate ainsi qu'au CaP sont donc toujours très étudiés et en développement. Ils présentent toutefois des limitations qui doivent être adressées dans les futures études.

1. Tout d'abord, la vérité terrain de la majorité des études est basée sur la biopsie, qui comporte plusieurs limitations :
   — **lésions non identifiées** : les lésions visées lors de biopsies guidées par IRM sont les lésions visibles à l'IRM. Un certain nombre de lésions non

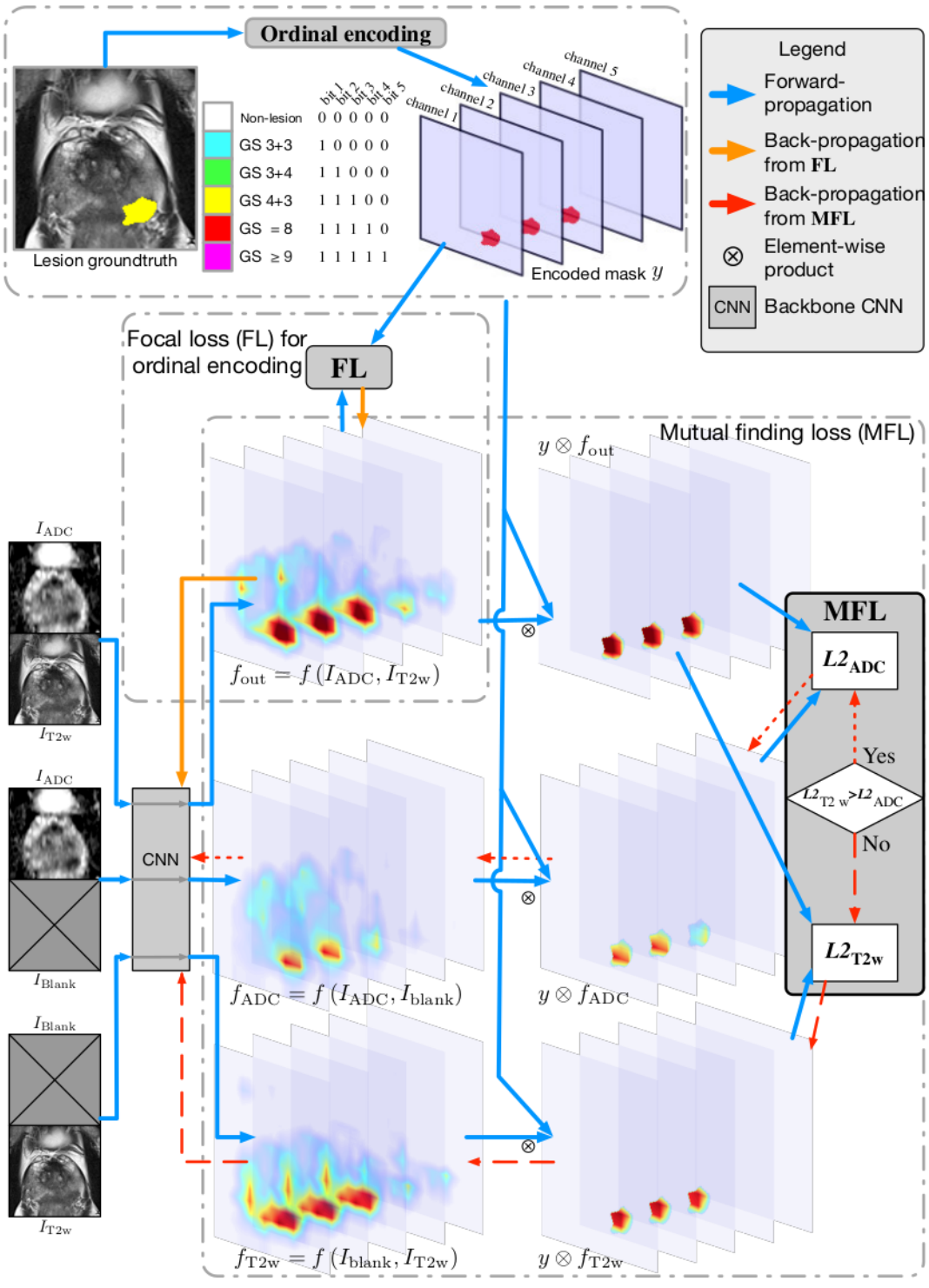


FIGURE 4.3 – Architecture du réseau FocalNet proposé par Cao et al. [15].

visibles à l'IRM ont de fortes chances de ne pas être identifiées si les biopsies systématiques ne tombent pas sur la lésion par hasard.

— **score de Gleason imprécis** : rappelons que la biopsie est un échantillonnage de la lésion et ne représente pas la lésion entière. Par exemple, l'échantillon de biopsie peut correspondre à une partie de la lésion où la majorité des cellules est de grade 4 et une minorité de grade 3, aboutissant à un GS 4+3 pour cette lésion. Or, il se peut que sur la majorité de la lésion se trouvent en réalité des cellules de grade 3 et une minorité de grade 4, ce qui correspondrait à un GS 3+4, moins agressif.

— **contours flous** : les contours de la lésion peuvent être flous à l'IRM et ne pas permettre une délimitation fiable de la lésion.

Pour toutes ces raisons, la référence pour caractériser des lésions avec fiabilité reste la pièce de prostatectomie. Cependant, cette vérité terrain s'obtient par l'ablation de la prostate et de ce fait est difficile à obtenir. C'est pourquoi la majorité des bases de données sont basées sur la biopsie.

2. Ensuite, les bases de données sont souvent homogènes et la généralisation des modèles n'est que rarement testée. La base ProstateX-2 peut permettre une évaluation des systèmes pour la détermination du GS mais n'est pas adaptée à un problème de segmentation. Une base de grande envergure pour la tâche de segmentation du CaP par agressivité n'existe pas à ce jour, et rend plus difficile une comparaison juste et sur des bases externes communes des différents modèles développés.
3. Enfin, très peu de travaux proposent une détermination du GS des lésions. Or, la caractérisation de l'agressivité des lésions pourrait permettre de diminuer les biopsies, douloureuses et invasives.

4.4 Conclusion

Ce chapitre nous a permis de faire un l'état de l'art des prototypes CAD développés dans le cadre du diagnostic du CaP par IRM avec des méthodes d'apprentissage profond supervisé. Au regard de cet état de l'art, nous avons proposé nos contributions, détaillées dans les chapitres suivants.

Papier	Base de données (#)	#	IRM	Séquences	Réseau	Loss	Dice prostate (%)	Dice ZP (%)	Dice ZT ou GC (%)	Commentaires
Wang et al., 2019 [129]	- privée (40) - PROMISE12 (50)	90	- Siemens TrioTim 3T (40) - variés (50)	T2-w axial	3D DSD-FCN	entropie croisée + similarité cosinus	- privée : 86 ± 4 - PROMISE12 : 88 ± 5	-	-	
Zabihollahy et al., 2019 [139]	privée	225	GE Discovery 3T	T2-w axial ou ADC	U-Net	dice loss	- T2 : 95.33 ± 7.77 - ADC : 92.09 ± 8.89	- T2 : 86.78 ± 3.72 - ADC : 86.1 ± 9.56	- T2 : 93.75 ± 8.91 - ADC : 89.89 ± 10.69	
Rundo et al., 2019 [108]	- privée (21) - I2CVB (19) - NCI-ISBI (40)	80	- Philips Achieva 3T (21) - Siemens TrioTim 3T (59)	T2-w axial	USE-Net	dice loss	-	- privée : 91.9 ± 2.1 - I2CVB : 83.1 ± 2.9 - NCI-ISBI : 80.1 ± 5.5	- privée : 87.1 ± 3.6 - I2CVB : 88.6 ± 1.5 - NCI-ISBI : 93.7 ± 1.0	Meilleures performances avec un modèle multi-source plutôt que source unique.
Ghavami et al., 2019 [50]	privée	232	Siemens Avanto 1.5T ou 3T	T2-w axial	VNet, Unet, HighRes3dNet, HolisticNet, Dense VNet, Adapted UNet	?	entre 86 et 90 selon les réseaux	-	-	
Zavala-Romero et al., 2020 [140]	- privée (220) - ProstateX (330)	550	- GE Discovery 3T (220) - Siemens 3T (330)	T2-w axial, sagital et coronal	multistream 3D U-Net	dice loss	- Siemens : 89.3±3.6 - GE : 82.5±11.2	- Siemens : 81.1 ± 7.9 - GE : 78.8 ± 9.3	-	Modèle combiné avec les 2 scanners meilleur que les modèles source unique.
Zhu et al., 2020 [146]	- privée (81) - PROMISE12 (50) - BWH (15)	146	- Philips 3T (81) - variés (65)	T2-w axial	BOWDA-Net	entropie croisée + distance + loss discriminateur	PROMISE12 : 92.54	-	-	
Liu et al., 2020 [86]	- ProstateX (304) - US (47)	351	- Siemens 3T	T2-w axial	Bayesian CNN	entropie croisée	-	- ProstateX : 80±5 - test externe : 79±6	- ProstateX : 89±4 - test externe : 87±7	Réseau meilleur que DeepLab V3+, Attention U-Net, R2U-Net, USE-Net et U-Net.
Liu et al., 2020 [85]	- NCI-ISBI (60) = sites A et B - I2CVB (19) = site C	79	- Siemens 3T (49) - Philips 1.5T (30)	T2-w axial	MS-Net (2D ResUNet)	dice loss + domain transfer loss	- site A : 91.54±2.01 - site B : 91.24±1.97 - site C :92.18±1.62	-	-	Réseau meilleur que USE-Net, Dual-Stream, Series-Adapter, Parallel-Adapter.
Cuocolo et al., 2021 [25]	ProstateX	204	Siemens 3T	T2-w axial	E-Net	Tversky loss	91± 4	71± 8	87 ± 5	Réseau meilleur que U-Net et ERFNet.
Jin et al., 2021 [68]	- PROMISE12 (100) - privée (106)	206	- variés (100) - ? (106)	T2-w axial	3D PBV-Net	dice loss	- PROMISE12 : 96.13 - privé : 97.65	-	-	
Isensee et al., 2021 [60]	- PROMISE12 (50) - Medical Segmentation Decathlon (32)	82	variés	- T2-w axial - T2-w axial et ADC	nnU-Net	entropie croisée + dice loss	PROMISE12 test : 91.94	Decathlon : 77	Decathlon : 90	
Bardis et al., 2021 [7]	privée	242	- Philips Ingenia 3T - Siemens Magnetom Trio 3T	T2-w axial	U-Net hybride 2D/3D	?	94.0 [93.0–96.1]	77.4 [72.7–83.2]	91.0 [89.4–93.8]	

TABLEAU 4.1 – État de l'art pour la segmentation des zones de la prostate sur des IRM avec des techniques d'apprentissage profond entre janvier 2019 et juillet 2021. Les papiers sont triés par date de publication croissante.

Papier	Base de données (#)	#	IRM	Séquences	Vérité terrain	Lésions incluses	Tâche	Réseau	Loss	Entraînement et évaluation	Critère pour vrai positif	Sensibilité	AUC, Dice (DSC) ou κ	Commentaires
Wang et al., 2018 [130]	- privée (156) - ProstateX (204)	360	- Siemens Skyra 3T (156) - Siemens 3T (204)	T2-w + ADC	biopsie	GG $\geq$ 2 et PI-RADS $\geq$ 4 pour ProstateX	détection des lésions	inspiré de GoogLeNet - co-entrainé	entropie croisée T2-w + entropie croisée ADC + inconsistency loss	5-fold cross-valid	point prédit dans la lésion	63.74% à 0.1FP 89.78% à 1FP	$AUC = 0.96$	AUC pour la tâche de classification et non de localisation.
Ishioka et al., 2018 [61]	privée	335	Philips Achieva 1.5T	T2-w	biopsie	GG $\geq$ 1	segmentation des lésions	Res U-Net	?	90% train 10% test (2 jeux de 17 patients)	pixels prédits dans la lésion	-	$AUC_{jeu1} = 0.645$ $AUC_{jeu2} = 0.636$	
Alkadi et al., 2018 [6]	I2CVB	19	Siemens TrioTim 3T	T2-w	biopsie	CaP (non précisé)	segmentation des lésions et de la prostate	VGG16	entropie croisée pondérée	60% train 10% val 25% test	?	-	$AUC = 0.995$	*Overfit* sur le peu de données disponibles.
Sumathipala et al., 2018 [117]	privée	186	?	T2W + ADC + DWI à haute valeur de b	biopsie et ou prostatectomie	CaP (non précisé)	?	Holistically nested edge detection	?	65% train 10% val 25% test	? pas papier		$AUC_{ZP} = 0.94 \pm 0.01$	6 institutions différentes.
Xu et al., 2019 [135]	ProstateX	346	Siemens 3T	T2-w + ADC + DWI à haute valeur de b	biopsie	PI-RADS $\geq$4	segmentation des lésions	ResNet	entropie croisée pondérée	70% train 30% val	intersection > 0	-	$AUC = 0.97$	
Schelb et al., 2019 [111]	privée	312	Siemens Prisma 3T	T2-w + ADC + DWI-B1500	biopsie	GG $\geq$ 2	segmentation des lésions et de la prostate	U-Net	entropie croisée	80% train : 4 réplicats de 4-fold cross-valid 20% test	à l'échelle du pixel ?	96% à 31% spéci 92% à 47% spéci	$DSC_{lesions} = 0.35$	Performances similaires au critère PI-RADS.
Dai et al., 2020 [26]	- privée (42) - ProstateX (78)	120	- Philips Ingenia 3T (42) - Siemens 3T (78)	T2-w + ADC	biopsie	? pour la base privée PI-RADS $\geq$ 4 pour ProstateX	segmentation des lésions et de la prostate	2D Mask R-CNN	?	2 expériences : - 58% ProstateX-2 train, 13% ProstateX-2 val, 100% privée et 29% ProstateX-2 test - idem pour ProstateX-2 mais 50% privée en train et 50% test	à l'échelle du pixel	- exp1 : 63% à 96% spéci ProstateX 22% à 97% spéci base privée - exp2 : 50% à 97% spéci ProstateX 33% à 97% spéci sur base privée	$DSC_{lesions} =$ - exp1 : 0.59 ProstateX 0.38 base privée - exp2 : 0.56 ProstateX 0.46 base privée	Dice sur lésions détectées seulement. Les performances sont améliorées pour leur base privée quand des patients sont inclus dans l'entraînement mais au détriment des performances ProstateX.
Chen et al., 2020 [22]	privée	136	- GE Discovery 3T (129) - GE Signa 3T (1) - Siemens Skyra 3T (6)	T2-w + ADC + DWI b1200	PI-RADS $\geq$4	PI-RADS $\geq$4	segmentation des lésions et de la prostate	MB-Net (Multi-Branch)	entropie croisée pondérée	60% train 20% val 20% test	à l'échelle du pixel	70.56% à 99.99% speci	$DSC_{cas} = 0.6333$ $DSC_{global} = 0.7205$	
Saha et al., 2021 [109]	privée	2436 (+296 jeu de test externe)	- Siemens Skyra ou Trim ou Prisma 3T (2436) - Siemens Skyra 3T (296)	T2-w + ADC + DWI à haute valeur de b	biopsie (et PI-RADS)	GG $\geq$ 2 parmi les PI-RADS $\geq$4	segmentation des lésions (prostate utilisée en entrée)	2 CNN 3D \\ : 1 3D U-Net seg ; 1 classifieur SERes-Net de patch 3D multicanal	loss focale + entropie croisée pondérée	65% train 15% val 20% test + 296 patients jeu externe	dice> 0.1	83.69 ± 5.22% sensi à 0.5FP 76.69% sur le jeu externe	$\kappa = 0.511$	Le kappa compare les prédictions du réseau à celles des radiologues et non à la vérité terrain.

TABLEAU 4.2 – État de l'art pour la détection ou segmentation binaire du cancer de la prostate sur des IRM avec des techniques d'apprentissage profond entre 2018 et juillet 2021. Les papiers sont triés par date de publication croissante.

Papier	Base de données (#)	#	IRM	Séquences	Vérité terrain	Lésions inclues	Réseau	Loss	Entraînement et évaluation	Sensibilité	AUC ou Kappa	Commentaires
Abraham et al., 2018 [3]	ProstateX-2 (99 train et 63 test)	162 (112 lésions train et 70 test)	Siemens 3T	T2-w + ADC + DWI à haute valeur b	biopsie	PI-RADS ≥ 4	SSAE à 3 couches	MSE + L2 reg	3-fold cross valid	80.26% pour prédire les CaP GG > 1	$\kappa_{train-set} = 0.2326$ $\kappa_{test-set} = 0.2772$	Entrée du réseau : 393 caractéristiques de texture.
Abraham et al., 2019 [2]	ProstateX-2 (train)	99 (112 lésions)	Siemens 3T	T2-w + ADC + DWI à haute valeur b	biopsie	PI-RADS ≥ 4	VGG16 + classifieur ordinal	?	LOPO cross-valid	-	$\kappa_{train-set} = 0.4727$ [0.2755, 0.6785]	
Chaddad et al., 2020 [16]	ProstateX-2 (train)	99 (112 lésions)	Siemens 3T	T2-w + ADC + DCE	biopsie	PI-RADS ≥ 4	NASNet pour extraire des caractéristiques puis forêt aléatoire	réseaux préentrainés sur ImageNet	5-fold cross-valid		$AUC_{GS6} = 80.08\%$ $AUC_{GS3+4} = 85.77\%$ $AUC_{GS4+3} = 97.30\%$ $AUC_{GS8} = 98.20\%$ $AUC_{GS\geq 9} = 86.51\%$	AUC : one versus all.

TABLEAU 4.3 – État de l'art pour la classification multiclasse (par PI-RADS ou GS) du cancer de la prostate sur des IRM avec des techniques d'apprentissage profond entre 2018 et juillet 2021. Les papiers sont triés par date de publication croissante.

Papier	Base de données (#)	#	IRM	Séquences	Vérité terrain	Lésions incluses	Tâche	Réseau	Loss	Entraînement et évaluation	Critère pour vrai positif	Sensibilité	Kappa ou dice	Commentaires
Cao et al., 2019 [15]	privée	417	Siemens 3T (4 modèles)	T2-w + ADC	prostatectomie	GG $\geq$1	détection des lésions par GG	FocalNet : DeepLab à 101 couches avec encodage ordinal	loss focale + Mutual finding loss	5-fold cross-valid	points dans un rayon de 5 mm des lésions	87.9% à 1FP pour lésions CS	-	Ne considèrent que les coupes avec au moins 1 lésion (donc biaise le nombre de faux positifs par patients dans les FROC).
Winkel et al., 2020 [132]	privée	2170 (8 sites) + 48 (1 site) test	Siemens Prisma 3T	T2-w + ADC + DWI-2000	biopsie	PI-RADS $\geq$ 3	détection des lésions par PI-RADS	réseau de localisation (CanLoc) + de qualification des candidats (CanQual) + de réduction des FP (FPR)	?	jeu de test externe	par cas : au moins une lésion PI-RADS $\geq$ 3 prédite chez le patient	87% à 50% speci PI-RADS 3 : 43% PI-RADS 4 : 73% PI-RADS 4 : 100%	$\kappa_{classif\_cas} = 0.42$	N'ont détecté que des lésions de la ZP. Calcul des sensibilités à l'échelle de la lésion pas clair.
De Vente et al., 2021 [27]	ProstateX-2 (99 train et 63 test)	162	Siemens 3T (2 modèles)	T2-w + ADC	biopsie	GG $\geq$2 et PI-RADS $\geq$ 4	segmentation des lésions par GG	U-Net 2D avec régression ordinale	entropie croisée pondérée	5-fold cross-valid +jeu de test ProstateX-2	toutes les lésions considérées comme vrai positif : si le centre de la lésion ne recoupe aucune prédiction, alors elle est classée GS 6	-	test : $DSC_{GG\geq 2} = 0.370 \pm 0.046$ $\kappa_{lesions} = 0.13 \pm 0.27$	

TABLEAU 4.4 – État de l'art pour la détection ou segmentation multiclasse (par PI-RADS ou GS) du cancer de la prostate sur des IRM avec des techniques d'apprentissage profond entre 2018 et juillet 2021. Les papiers sont triés par date de publication croissante.

Chapitre 5

Base de données CLARA-P

Sommaire

5.1 Introduction

La qualité d'une base de données est primordiale à l'établissement d'un modèle statistique. En particulier, l'entraînement d'un modèle de segmentation dépend directement des images et annotations utilisées lors de l'apprentissage. L'acquisition des données nécessite une expertise médicale ou une collaboration avec un centre de santé. L'inclusion des images des patients dans une base de données impose l'accord de ces derniers, les données médicales étant des données sensibles. Une fois l'autorisation obtenue, les données doivent être anonymisées. Enfin, le plus chronophage reste la création de la vérité terrain. Pour une base de données ayant pour vocation de servir à des tâches de segmentation, une annotation fine à l'échelle du pixel est requise. Elle nécessite potentiellement des croisements avec le gold-standard si la vérité ne peut pas être connue par simple analyse de l'image, comme c'est le cas pour le cancer de la prostate avec l'analyse de la pièce de prostatectomie ou l'échantillon de biopsie à confronter aux IRM. Un temps expert médical important est nécessaire pour ensuite annoter chacun des volumes 3D. La nature répétitive et la complexité du protocole d'annotation rendent la reproduction de cette tâche difficile à l'échelle d'une grande base d'imagerie. À ce titre, on observe couramment une forte variabilité entre les annotations de plusieurs experts (variabilité inter-observateur), ou lorsqu'un même expert analyse la même image (variabilité intra-observateur). Enfin, le transfert et le stockage des données est également réglementé et requiert de suivre des procédures exigeantes. La complexité liée à la création d'une base de données médicales a souvent pour conséquence de limiter le nombre de patients présents

dans les bases. En outre, peu de bases de données publiques existent pour certaines applications médicales, rendant plus difficile l'apprentissage de modèles robustes.

Le travail présenté dans cette thèse exploite une base de données utilisée pour l'entraînement et la validation des modèles, la base CLARA-P, pour **C**orré**L**ations **A**natomo-**RA**diologiques en IRM de **P**rostate. La constitution de cette base a commencé en septembre 2008, à l'initiative du Pr. Olivier Rouvière à Lyon, après déclaration aux autorités administratives compétentes (*Comité de Protection des Personnes, référence L 09-04* et *Commission Nationale de l'Informatique et des Libertés, traitement n° 08-06*). Elle est notamment décrite dans Bratan et al. [10]. D'autres bases de données sont élaborées par le Pr. Olivier Rouvière, dont la base CLARA-B, pour laquelle la vérité est cette-fois ci la biopsie.

Nous avons travaillé à partir des 290 patients inclus dans la base CLARA-P jusqu'en 2014. Tous ont subi une prostatectomie radicale (voir section 2.6) après que la présence de cancer a été avérée par biopsies. Les spécimens de prostatectomie ont été analysés *a posteriori* par un anatomo-pathologiste (10 ans d'expérience à la création de la base) en tenant compte des directives internationales [110], fournissant ainsi la vérité terrain histologique. Après corrélation avec les coupes histologiques, les uroradiologues ont reporté en consensus les différentes lésions.

Ce sont les IRM-mp associées à la vérité terrain tracée par les uroradiologues qui sont utilisées pour l'entraînement des modèles présentés chapitre 7, 8 et 9.

Ce chapitre décrit cette base de données CLARA-P, en donne des statistiques et présente les exclusions réalisées.

5.2 Acquisition des données IRM multiparamétriques (IRM-mp)

Les examens IRM-mp ont été effectués selon un protocole standardisé, sur des scanners cliniques. 4 scanners différents ont été utilisés pour l'acquisition des images, au sein de trois départements de radiologie des Hospices Civils de Lyon (HCL) :

— Siemens Magnetom Symphony 1.5T à l'hôpital Edouard Herriot à Lyon

— GE Discovery 3T à l'hôpital Edouard Herriot à Lyon

— Philips Achieva 3T à l'hôpital Pierre Wertheimer à Bron

— Philips Ingenia 3T à l'hôpital Lyon Sud à Pierre-Bénite

Il n'y a pas eu de sélection quant à la répartition des patients sur les différentes IRM (présentée tableau 5.2). Chaque acquisition multiparamétrique comprend les séquences T2-w, DWI et DCE (présentées section 2.7). Les paramètres utilisés pour les différents scanners sont présentés tableau 5.1. À noter qu'une antenne pelvienne (voir figure 2.7A) a été utilisée pour toutes les acquisitions, mais associée à une antenne endorectale (voir figure 2.7B) pour certains patients acquis sur le scanner Philips Achieva seulement (43/55 patients).

Pour l'imagerie DCE, une injection intraveineuse de 0,2 mL/kg de gadotérate méglumine (Dotarem ; Guerbet, Roissy, France) a été réalisée à 3 mL/s dans tous les cas. La résolution temporelle a été adaptée à l'intensité du champ et à la configuration de la bobine.

Les images axiales T2-w, DWI et DCE ont été acquises avec la même épaisseur de coupe et la même position afin de permettre une comparaison directe entre les différentes séquences.

Scanner	Champ	Séquence	T_R (ms)	T_E (ms)	FOV (mm)	Matrice (voxels)	Dimension voxel (mm)	Valeurs de b (s/mm^2)
Siemens Symphony	1.5T	T2w	7750	109	200×200	256×256	$.78 \times .78 \times 3$	-
		DWI	4800	90	300×206	128×88	$2.34 \times 2.34 \times 3$	0, 600
		DCE	5.38	2.73	210×240	448×512	$.47 \times .47 \times 3$	-
GE Discovery	3T	T2w	5000	104	220×220	512×512	$.43 \times .43 \times 3$	-
		DWI	5000	90	380×380	256×256	$1.48 \times 1.48 \times 3$	0, 2000
		DCE	3.9	1.7	240×192	180×160	$.94 \times .94 \times 3$	-
Philips Ingenia	3T	T2w	~ 5500	100	180×180	336×336	$.54 \times .54 \times 3$	-
		DWI	~ 4800	~ 81	350×350	288×288	$1.22 \times 1.22 \times 3$	2000
		DCE	3.92	2.30	160×160	160×160	$1.00 \times 1.00 \times 3$	-
Philips Achieva	3T	T2w	5021	120	180×180	352×352	$.54 \times .54 \times 3$	-
		DWI	3925	70	180×180	176×176	$1.03 \times 1.03 \times 3$	0, 800, 2000
		DCE	4	2.3	180×180	176×176	$1.02 \times 1.02 \times 3$	-

TABLEAU 5.1 – Paramètres utilisés pour l'imagerie de la prostate sur les différents scanners de la base CLARA-P. Lorsque plusieurs valeurs ont été utilisées pour un scanner et une modalité donnés, la valeur la plus fréquente est reportée. T_R : temps de répétition, T_E : temps d'écho, FOV : champ de vue (de l'anglais *Field Of View*)

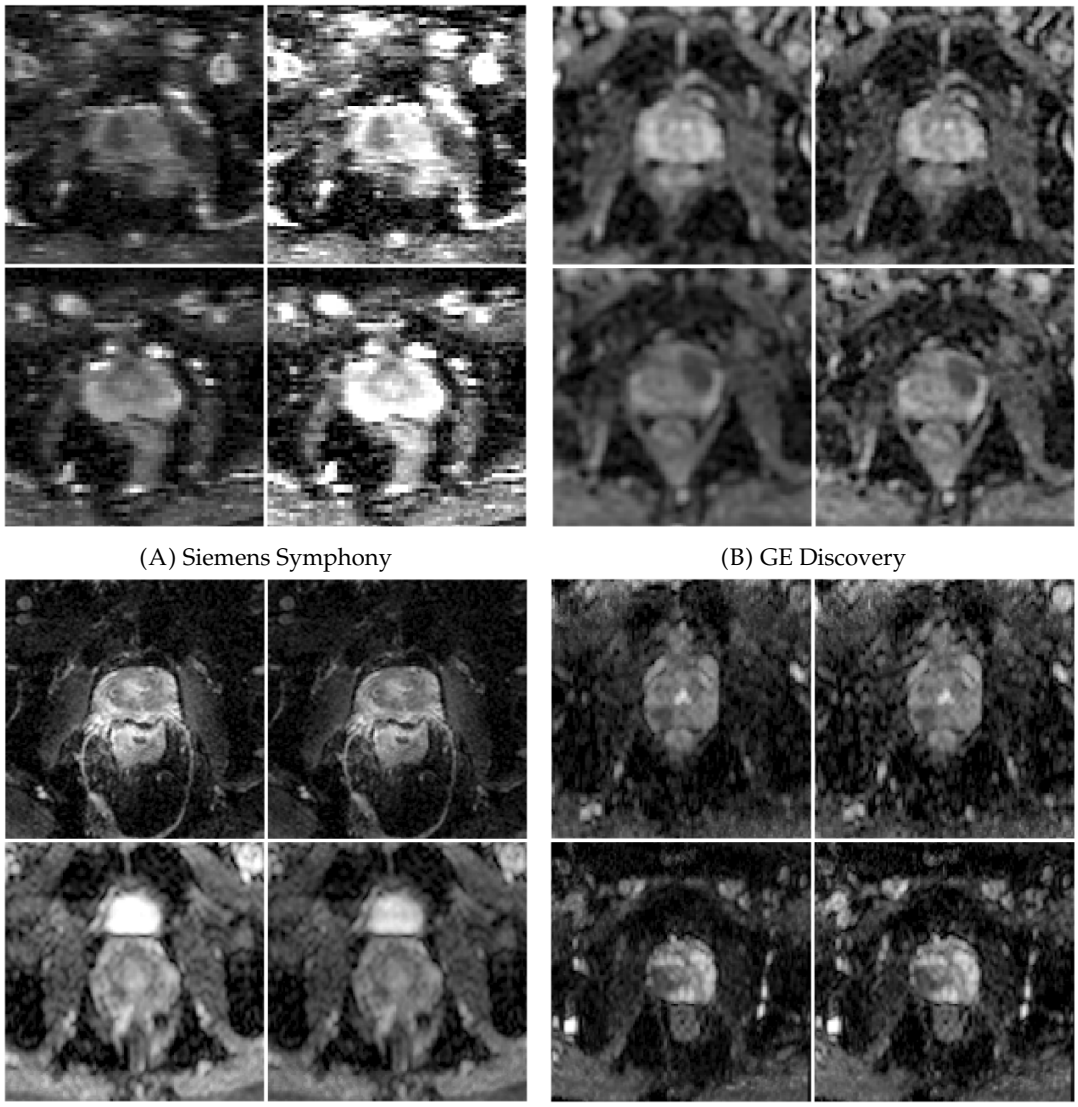
Calcul des cartes ADC

Les cartes ADC sont calculées à partir de plusieurs volumes de diffusion, acquis à différentes valeurs de b (voir section 2.7.3). Les scanners produisent automatiquement ces cartes paramétriques mais les méthodes utilisées varient selon les constructeurs. En outre, les volumes ADC n'ont pas pu être retrouvés pour certains patients. Pour des raisons d'homogénéité et pour résoudre le problème des informations manquantes, les cartes ont été recalculées *a posteriori* sur MATLAB par Tristan Jaouen (doctorant au LabTAU) dans le cadre du projet PERFUSE. Elles ont été obtenues par ajustement avec la méthode des moindres carrés de la courbe en échelle logarithmique des intensités en fonction de la valeur de b (2 ou 5 volumes acquis à différentes valeurs de b disponibles selon les patients). Des exemples de cartes ADC brutes et recalculées sont donnés figure 5.1 pour les différents scanners de la base.

5.3 Analyse des images IRM

Les IRMs des patients inclus dans la base CLARA-P ont été revus indépendamment par deux uroradiologues de 11 ans et 1 an d'expérience à la création de la base en 2008. Les lecteurs n'avaient pas connaissance des données cliniques, biologiques ou histopathologiques des patients, mais savaient qu'ils étaient concernés par une prostatectomie.

Tout d'abord, les uroradiologues ont noté et localisé sur une cartographie de la prostate en 27 secteurs (voir figure 2.13) toutes les anormalités visibles. Dans la zone périphérique (ZP), toutes les anormalités avec un signal de faible intensité sur le T2-w et/ou sur les cartes ADC et/ou avec un rehaussement précoce sur les images DCE ont été considérées. Dans la zone de transition (ZT), seulement des aires homogènes avec un faible signal sur le T2-w, des marges floues, sans capsule visible et sans composante kystique ont été notées comme suspectes. Les cartes ADC et les images de perfusion ont été interprétées visuellement seulement, aucune valeur quantitative n'a été utilisée pour diagnostiquer le cancer.



(A) Siemens Symphony (B) GE Discovery

(C) Philips Achieva (D) Philips Ingenia

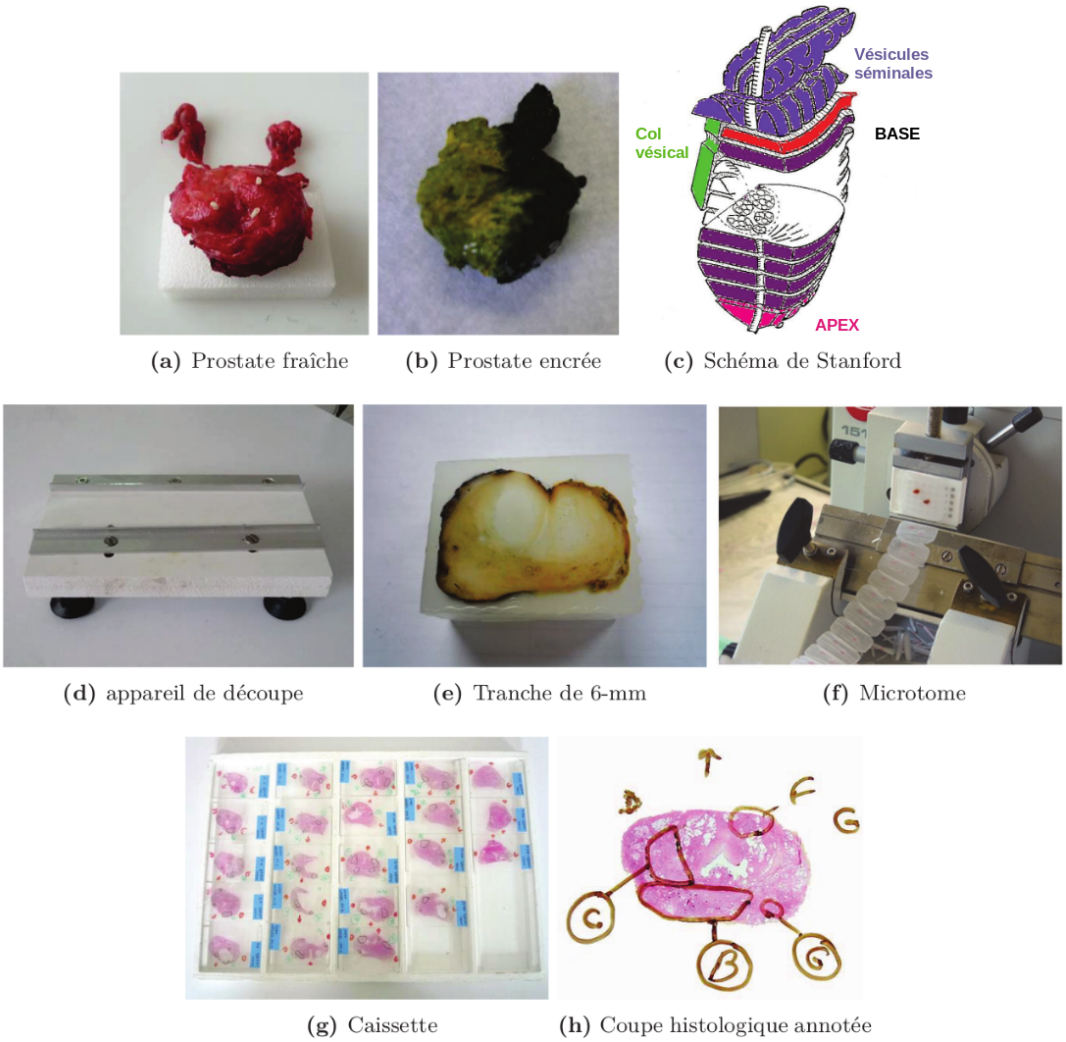
FIGURE 5.1 – Exemple de cartes ADC pour les 4 scanners de la base CLARA-P. Dans chaque bloc de 4 images, la 1ère colonne correspond à la carte brute calculée par le constructeur et la 2ème à la carte ADC recalculée. Chaque ligne montre un patient différent.

Un score correspondant au degré de suspicion de malignité a été attribué à chacune des régions identifiées :

— 0 : bénignité certaine
— 1 : probablement bénin
— 2 : intermédiaire
— 3 : probablement malin
— 4 : malignité certaine

5.4 Analyse des données histologiques

5.4.1 Préparation des pièces de prostatectomie



(a) Prostate fraîche (b) Prostate encrée (c) Schéma de Stanford

(d) appareil de découpe (e) Tranche de 6-mm (f) Microtome

(g) Caissette (h) Coupe histologique annotée

FIGURE 5.2 – Différentes étapes de la préparation de la pièce de prostatectomie. Adaptée de Niaf [94].

Les pièces de prostatectomie de chacun des patients ont été traitées selon les étapes décrites ci-dessous :

1. La pièce est encrée avec plusieurs couleurs : jaune pour le lobe droit, noire pour le lobe gauche et rouge pour la face antérieure (figure 5.2b).

2. La pièce est fixée pendant 24 heures minimum en solution de formaldéhyde.
3. L'apex, le col vésical et la base sont isolés et coupés de façon sagittale et parasagittale (conisation) pour l'analyse des marges (voir figure 5.2c).
4. Le reste de la glande (en violet sur la figure 5.2c) est coupé de l'apex à la base dans un plan axial en tronçons de 6 mm d'épaisseur (figure 5.2e) avec une machine spécialement conçue en interne (figure 5.2d) pour garantir que les tranches soient uniformément espacées.
5. Les tranches sont mises dans des cassettes et placées dans du formaldéhyde pour une nouvelle période de fixation de 24 à 48 heures. Elles sont ensuite traitées et incluses en paraffine.
6. Les blocs de paraffine ainsi obtenus sont coupés au microtome (figure 5.2f) tous les 0.5-1.5 mm.
7. Enfin, les rubans obtenus sont étalés sur lames et colorés à l'Hématoxyline-Eosine Safran (voir figure 5.2g).

5.4.2 Analyse des coupes histologiques

L'analyse au microscope de l'ensemble des coupes a été réalisée par une anatomo-pathologiste (10 ans d'expérience à la création de la base en 2008), qui n'avait pas connaissance des données IRM.

Toutes les tumeurs ayant une surface dans le plan supérieure ou égale à 2 mm$\times$ 2 mm, visibles sur deux coupes consécutives et de GS $\geq$ 5 ont été délimitées sur chacune des coupes ; les autres n'étant pas considérées. Les régions malignes séparées de moins de 1 mm l'une de l'autre dans le même plan, avec la même architecture et le même GS ont été considérées comme faisant partie de la même tumeur.

Enfin, des coupes ayant 3 mm d'espacement ont été sélectionnées pour former un jeu de coupes pour la corrélation entre l'IRM et la pathologie (voir figure 5.2h).

5.5 Corrélation anatomo-radiologique

Les analyses histologiques (voir section 5.4.2) et IRM (voir section 5.3) ont ensuite été confrontées par les uroradiologues et l'anatomo-pathologiste selon la procédure suivante :

1. Les uroradiologues révèlent les zones suspectes identifiées
2. Toutes les zones malignes repérées sur les coupes histologiques sont reportées sur les images IRM-mp à l'aide d'un maximum de repères (par exemple kystes, nodules d'hyperplasie, canaux éjaculatoires).
3. L'anatomo-pathologiste décide pour chacune des zones si elle correspond à un cancer identifié histologiquement ou pas
   — une correspondance est considérée comme un vrai positif (VP) si le plus large diamètre est compris dans $50-150\%$ du plus large diamètre du cancer histologique correspondant ;
   — sinon, l'anormalité repérée sur l'IRM est considérée comme un faux positif (FP) et le cancer faux négatif (FN).

La segmentation manuelle a été réalisée par les radiologues en utilisant la station de visualisation d'images OsiriX® (Genève, Suisse). On note que, suivant la référence histologique, la lésion entière est contourée et pas seulement la région de forte anormalité (pic de rehaussement sur la DCE par exemple).

5.6 Contourage des zones anatomiques

En plus des lésions, les zones anatomiques (ZP et ZT) ont été délimitées sur chacune des séquences par différents experts (radiologues, interne en médecine ou doctorant) selon les patients. Le logiciel OsiriX® est également utilisé.

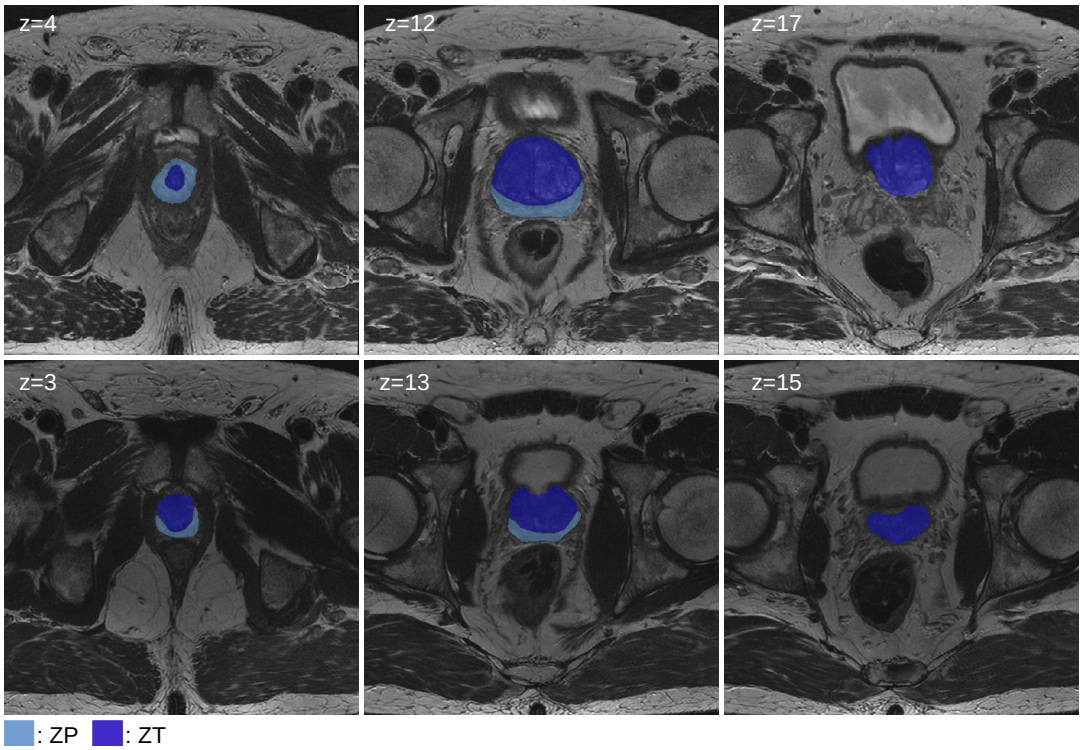


FIGURE 5.3 – Exemples de contours des zones anatomiques ZP et ZT sur les séquences T2-w pour plusieurs coupes localisées à l'apex (z faible), au milieu et à la base (z élevé).

5.7 Nettoyage de la base de données

Au cours des 7 années d'inclusion dans la base CLARA-P, 290 patients ont été enregistrés.

5.7.1 Patients exclus

Parmi les patients enregistrés, plusieurs ont été exclus par les radiologues pour diverses raisons :

— 2 car au moins une séquence était manquante

— 1 car les contours n'avaient pas été faits et les photos des lames non disponibles pour tracer de nouveau les contours

— 6 car l'IRM était de trop mauvaise qualité pour reporter correctement les lésions

— 1 à cause de présence de sang dans l'IRM, réalisé après biopsie

— 2 à cause d'artefacts dus au port d'une prothèse

— 1 car la durée d'acquisition de la perfusion était trop courte

L'exclusion de ces 13 patients a amené à 277 le nombre de patients dans la base.

J'ai également exclu des patients de mon côté :
— 1 car les contours des lésions n'étaient pas disponibles
— 1 car l'ADC était de trop mauvaise qualité

Après les exclusions, les données de 275 patients sont exploitables.

5.7.2 Vérité terrain

Un premier travail de vérification et de mise à jour de la base de données a été effectué au début de la thèse. Ce travail a été réalisé en étroite collaboration avec Tristan Jaouen, doctorant encadré par le Pr. Olivier Rouvière et Rémi Souchon au laboratoire LabTAU dans le cadre du projet PERFUSE. Pour cette vérification, nous nous sommes basé sur un fichier Excel servant de référence. Il contient, pour chaque patient, les cancers annotés à l'aveugle par chacun des deux radiologues et la vérité anatomo-pathologique correspondante.

Il a s'agit de :
— vérifier que toutes les séquences IRM étaient disponibles pour chacun des patients, les retrouver et compléter la base le cas échéant, exclure le patient sinon ;
— retrouver et regrouper les contours qui avaient été dessinés depuis le commencement de la base en 2008 ;
— vérifier que pour chacune des lésions reportées dans le fichier Excel, un contour sur l'IRM avait été fait, et ce, sur chacune des modalités ;
— vérifier que les masques étaient bien superposables et cohérents avec l'IRM du patient. Dans certains cas, le nombre de coupes dans le masque et l'IRM différaient, ou l'ordre des coupes était inversé. Dans d'autres, les contours étaient manquants sur certaines coupes. Dans certains cas, la lésion était notée comme en ZP par exemple mais correspondait visuellement à une lésion dans la ZT ;
— identifier les contours de mauvaise qualité et les communiquer au radiologue (pour les lésions) ou à l'interne (pour les contours de prostate) pour les refaire ;
— corriger des erreurs dans le fichier Excel servant de référence : erreur au niveau du score de Gleason indiqué, de la localisation de la lésion (ZP ou ZT), ou encore lésion notée comme "visible" mais impossible à reporter sur l'IRM.

Tristan Jaouen a également travaillé à l'anonymisation des données et à leur transfert sur un serveur XNAT, en collaboration avec Sylvain Gouttard, ingénieur aux Hospices Civiles de Lyon.

Ce travail de mise au propre de la base, jusqu'à la mise en production sur XNAT, a nécessité plus d'un an de travail.

La base de données stockée sur XNAT contient, pour chaque patient :
— les fichiers `.dicom` des différentes séquences IRM ;
— des photos `.png` des lames anatomo-pathologiques ;
— les contours des lésions et zones anatomiques (ZP et ZT) sous différents formats :
  — `.rois_series` (format propre à OsiriX®)
  — `.xml` (exportés depuis OsiriX®)
  — `.mat` (obtenus après conversion des fichiers `.xml` avec MATLAB)

5.8 Composition de la base de données

Par la suite, nous considérons la base de données après exclusion des données non exploitables et donc composée de 275 données patients.

Informations cliniques La répartition des patients par scanner est présentée tableau 5.2, associée au taux de PSA et âge moyen des patients. Les patients ont une soixantaine d'années en moyenne et un taux moyen de PSA proche de 9, mais avec un écart type important (± 8.33).

Scanner	# Patients	Âge moyen	PSA moyen
Siemens Symphony	68	60.85 ± 5.89	9.63 ± 7.31
GE Discovery	126	62.29 ± 5.75	8.62 ± 9.22
Philips Achieva	55	61.04 ± 4.76	8.43 ± 8.31
Philips Ingenia	26	63.77 ± 4.00	8.50 ± 6.38
Total	275	61.82 ± 5.51	8.82 ± 8.33

TABLEAU 5.2 – Statistiques cliniques et répartition par scanner des 275 patients de la base CLARA-P.

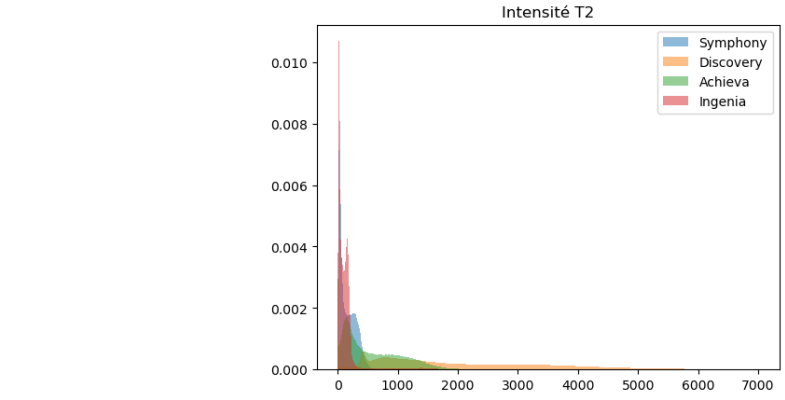
Images IRM Les intensités en niveaux de gris des données IRM sont présentées figure 5.4 pour le T2-w et figure 5.5 pour l'ADC. Pour le T2-w, la distribution des intensités est très variable selon les scanners : les intensités maximales des scanners Siemens Symphony et Philips Ingenia sont inférieures à 1000 alors qu'elles dépassent 2000 pour Philips Achieva et sont plus étalées pour GE Discovery, avec des intensités en niveaux de gris supérieures à 6000.

Concernant les cartes paramétriques ADC, les intensités sont dans le même intervalle quel que soit le scanner, de par la manière dont sont calculées de ces cartes paramétriques.

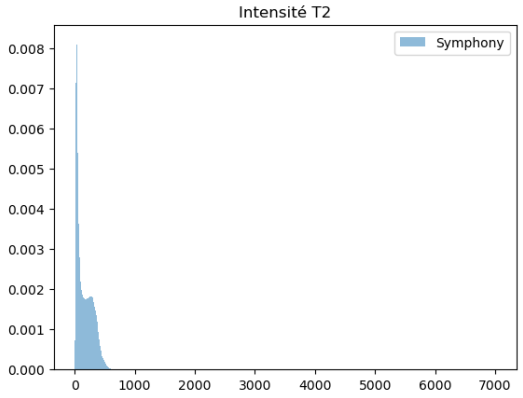
Annotations des lésions Les patients ayant subi une prostatectomie, la base contient essentiellement des patients ayant un cancer cliniquement significatif (noté CS CaP). En effet, 215/275 patients ont un cancer de grade GS $\geq 3+4$, soit plus de 78 % des patients.

Le nombre de lésions par scanner et par classe est présenté tableau 5.3, avec les détails par zone (ZP ou ZT) tableau 5.4 et tableau 5.5. Le tableau 5.6 synthétise le nombre de lésions dans la base de 275 patients, sans considérer le scanner de provenance. À noter qu'une lésion est ici définie comme un groupe de voxels de même classe de taille $\geq$ 45mm^3. Le tableau 5.7 présente des statistiques concernant les tailles des lésions

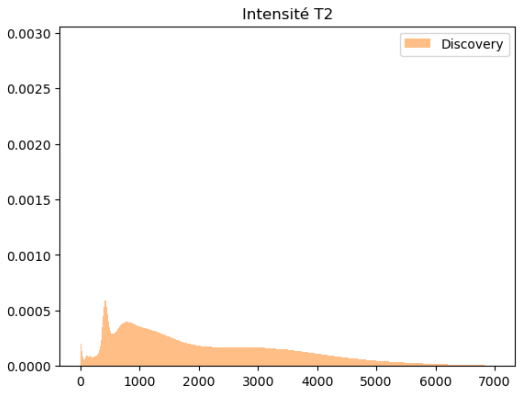
Performances des radiologues Elles sont calculées à partir du fichier Excel contenant les régions suspectes identifiées par les radiologues. Chaque région a été définie comme VP, FP, ou FN après corrélation avec les pièces histologies. Pour chaque GS, la sensibilité est calculée en comptant le nombre de lésions identifiées parmi toutes les lésions de cette agressivité. Le nombre de FP ne peut pas être calculé par classe puisque les radiologues ne caractérisent pas le GS de chaque zone suspecte. Il correspond donc au nombre moyen de région FP par patients.



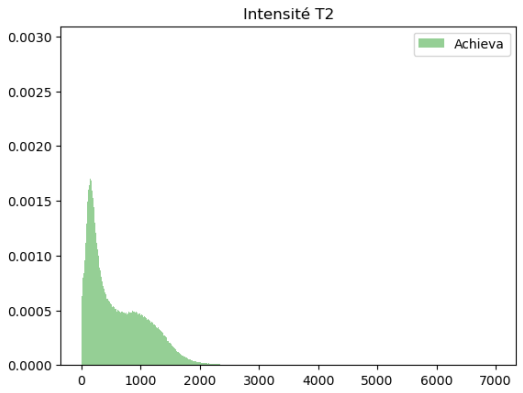
(A) Tous scanners



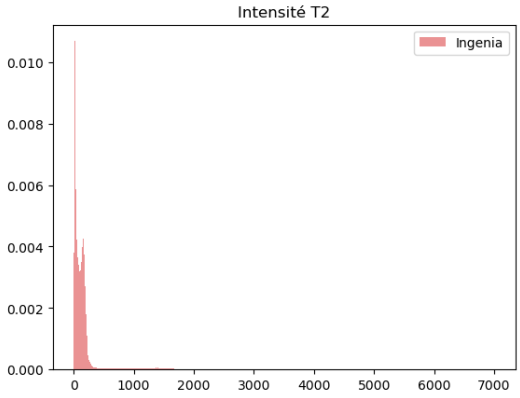
(B) Siemens Symphony



(C) GE Discovery

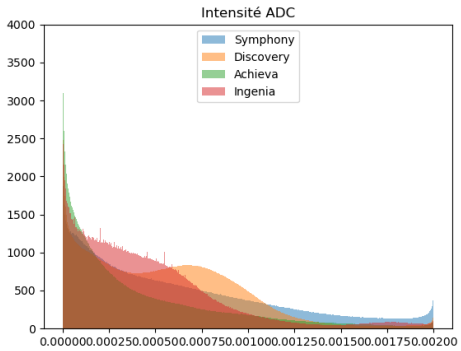


(D) Philips Achieva

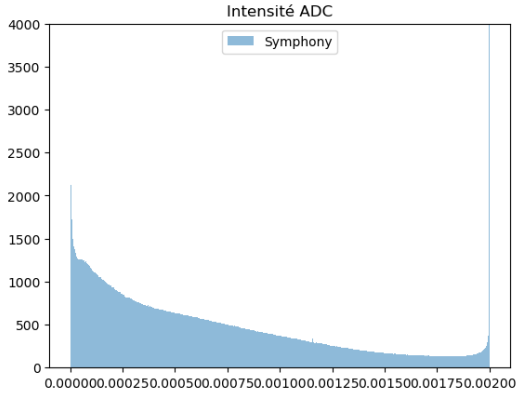


(E) Philips Ingenia

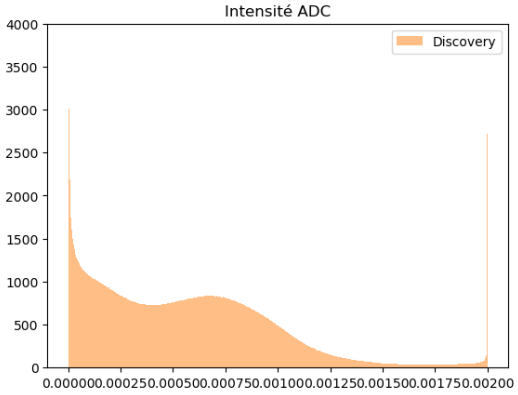
FIGURE 5.4 – Distribution des intensités des voxels composant les volumes T2-w par scanner.



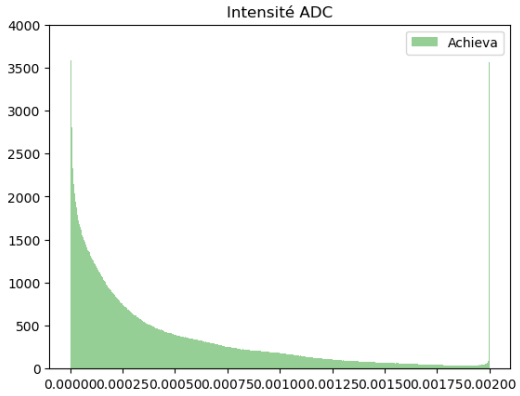
(A) Tous scanners



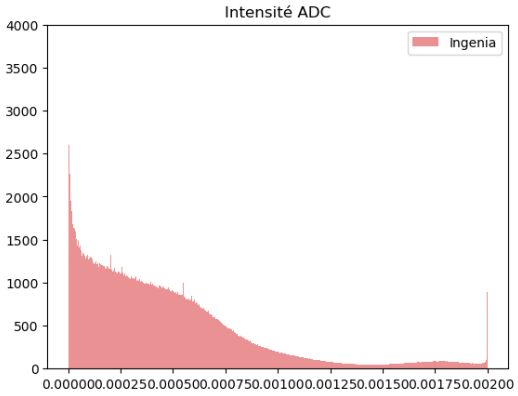
(B) Siemens Symphony



(C) GE Discovery

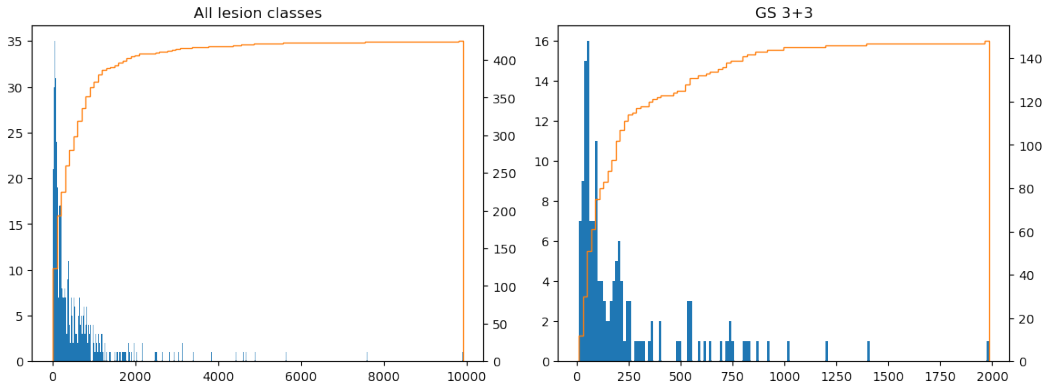


(D) Philips Achieva



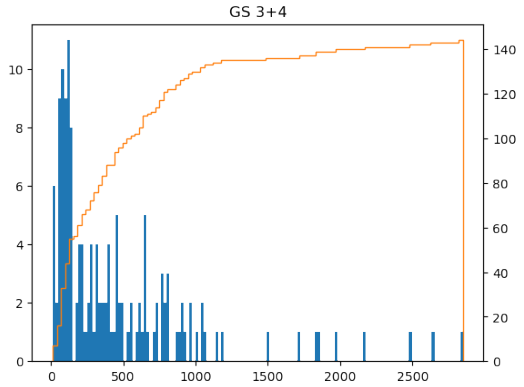
(E) Philips Ingenia

FIGURE 5.5 – Distribution des intensités des voxels composant les cartes paramétriques ADC par scanner.

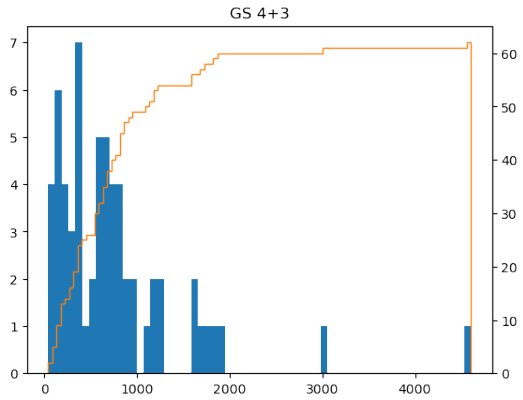


(A) tous GS

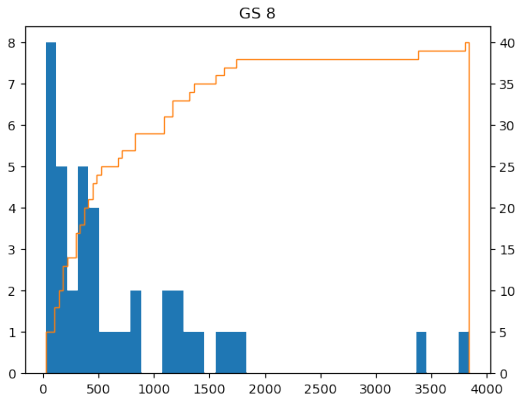
(B) GS 3+3



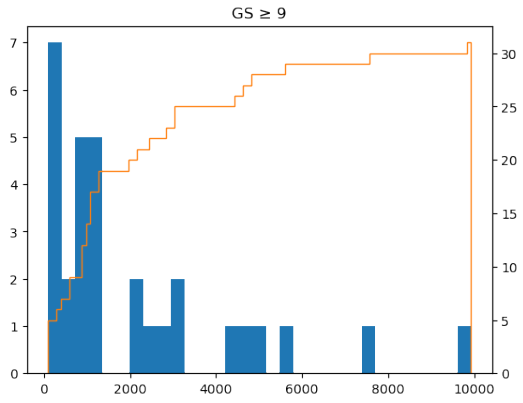
(C) GS 3+4



(D) GS 4+3



(E) GS 8



(F) GS $\geq$ 9

FIGURE 5.6 – Histogramme (bleu) et histogramme cumulé (orange) des tailles de lésion 3D en voxels (A) toutes classes confondues et (B :F) par GS. Les images ont préalablement été ré-échantillonnées à une résolution de $1 \times 1 \times 3$ mm^3.

Scanner	GS 3+3	GS 3+4	GS 4+3	GS 8	GS $\geq$ 9	Total
Siemens Symphony	47	40	12	11	6	116
GE Discovery	53	64	33	16	14	180
Philips Achieva	38	16	6	10	9	79
Philips Ingenia	7	23	11	3	2	46
Total	145	143	62	40	31	421

TABLEAU 5.3 – Distribution des lésions par scanner et score de Gleason (GS) pour les 275 patients de la base CLARA-P. Une lésion est définie comme un groupe de voxels de même classe de taille $\geq 45\text{mm}^3$.

Scanner	GS 3+3	GS 3+4	GS 4+3	GS 8	GS $\geq$ 9	Total
Siemens Symphony	39	35	11	7	5	97
GE Discovery	42	58	30	14	12	156
Philips Achieva	31	15	5	9	8	68
Philips Ingenia	5	19	11	3	2	40
Total	117	127	57	33	27	361

TABLEAU 5.4 – Distribution des lésions de la ZP par scanner et score de Gleason (GS) pour les 275 patients de la base CLARA-P.

Scanner	GS 3+3	GS 3+4	GS 4+3	GS 8	GS $\geq$ 9	Total
Siemens Symphony	8	5	1	4	1	19
GE Discovery	11	6	3	2	2	24
Philips Achieva	7	1	1	1	1	11
Philips Ingenia	2	4	0	0	0	6
Total	28	16	5	7	4	60

TABLEAU 5.5 – Distribution des lésions de la ZT par scanner et score de Gleason (GS) pour les 275 patients de la base CLARA-P.

	GS 3+3	GS 3+4	GS 4+3	GS 8	GS $\geq$ 9	Total
ZP	117	127	57	33	27	361
ZT	28	16	5	7	4	60
Total	145	143	62	40	31	421

TABLEAU 5.6 – Distribution des lésions par score de Gleason (GS) pour les 275 patients de la base CLARA-P.

	GS 3+3	GS 3+4	GS 4+3	GS 8	GS $\geq$ 9	Tous GS
#	148	144	62	40	31	425
taille min	9	12	37	27	97	9
taille max	1 985	2 850	4 603	3 839	9 918	9 918
taille moyenne	231	470	760	718	2 143	574
taille médiane	106	292	602	407	1 121	260
# taille $<$ 15	3	1	0	0	0	4
# voxels	34 242	67 712	47 174	28 737	66 461	244 326

TABLEAU 5.7 – Statistiques concernant la distribution des tailles de lésion 3D en voxels par score de Gleason (GS) pour les 275 patients de la base CLARA-P. Les images ont préalablement été ré-échantillonnées à une résolution de $1 \times 1 \times 3$ mm^3.

	GS 3+3		GS 3+4		GS 4+3		GS 8		GS $\geq$ 9		Tous GS		FP	
	OR	FB	OR	FB	OR	FB	OR	FB	OR	FB	OR	FB	OR	FB
Siemens	0.63	0.39	0.88	0.83	0.91	0.91	1.0	1.0	1.0	1.0	0.80	0.70	0.68	0.60
GE	0.63	0.61	0.89	0.79	0.93	0.90	0.94	0.87	0.93	0.93	0.83	0.78	0.64	0.50
Achieva	0.69	0.63	1.0	0.94	0.83	0.83	0.87	0.87	0.86	0.86	0.81	0.76	1.18	1.04
Ingenia	0.71	0.57	0.64	0.68	1.0	0.90	1.0	1.0	1.0	1.0	0.77	0.75	0.35	0.27
Total	0.65	0.54	0.86	0.80	0.93	0.90	0.95	0.92	0.93	0.93	0.81	0.75	0.73	0.61

TABLEAU 5.8 – Performance des 2 uroradiologues (OR et FB) pour les 275 patients de la base CLARA-P. Sont rapportés les sensibilités par score de Gleason GS et le nombre moyen de Faux Positif (FP) par patient, sachant que la tâche des radiologues consiste à repérer les lésions CS sans leur assigner de GS. Ici, seules les lésions visibles *a posteriori* sont considérées.

5.9 Conclusion

La base de données CLARA-P est une base de grande qualité, avec la vérité terrain la plus fiable qu'il soit, obtenue après prostatectomie. En conséquence, la base est également biaisée, avec une prévalence de patients ayant au moins un cancer cliniquement significatif (215/275 patients).

La base présente une grande hétérogénéité, de par les différents scanners utilisés, correspondant à différents modèles et constructeurs, à différents champs (1.5T et 3T), avec l'utilisation ou non de sonde endorectale, mais aussi à des paramètres d'acquisition différents, qui ont évolué en accord avec les recommandations entre 2008 et 2014, date de la dernière acquisition.

Cette base CLARA-P a été au cœur des développements méthodologiques que nous présentons dans les chapitres suivants.

Chapitre 6

Choix méthodologiques

Sommaire

6.1 Introduction

Dans ce chapitre, nous présentons les choix méthodologiques que nous avons faits pour répondre à notre objectif scientifique, qui est de construire un CAD multiclasse robuste à des donnés hétérogènes. Dans une première partie, nous positionnons nos contributions par rapport à l'état de l'art. Dans une seconde partie, nous présentons les choix concernant les données et le *pipeline* d'évaluation, commun à toutes les contributions. Ils s'appliquent aux trois chapitres d'expériences suivants (chapitre 7, chapitre 8 et chapitre 9).

6.2 Positionnement des contributions par rapport à l'état de l'art

L'état de l'art présenté chapitre 4 atteste du nombre important de systèmes CAD étudiés pour la caractérisation (CADx) ou la détection (CADe) du CaP à partir d'images IRM. Toutefois, la grande majorité des modèles cherche à résoudre un problème de classification ou de segmentation binaire et non multiclasse. Pour répondre à notre objectif, c'est-à-dire construire un CAD multiclasse, nous avons abordé cette problématique en plusieurs étapes :

— Tout d'abord, nous avons étudié une approche totalement supervisée, exploitant les données annotées à disposition. Nous avons décidé de développer directement un système CADe, c'est-à-dire qui localise les lésions dans les images. En effet, la faisabilité de tels systèmes a déjà été démontrée dans des études dans le cas binaire (voir section 4.3.2) et notre but final est de localiser et

non classer les lésions (systèmes CADx). Pour choisir le modèle de segmentation de base, nous avons fait une évaluation comparative de différents réseaux. Nous avons sélectionné le réseau U-Net (présenté section 3.4.1), qui, avec ses variantes, a fait ses preuves dans diverses applications médicales et également dans l'IRM de prostate [61, 139, 111, 108, 140, 85, 7]. Nous avons ensuite cherché comment améliorer les performances du modèle et exploré la piste des modèles d'attention. Les contours de la prostate étant disponible dans la base, nous avons cherché à exploiter cette information précieuse pour concentrer l'attention du réseau sur la prostate, comme le ferait un radiologue. Cette partie correspond au chapitre 7.

— Ce premier modèle supervisé est limitant pour l'inclusion de données supplémentaires par son besoin de données totalement annotées. En effet, pour améliorer les modèles et leur généralisation, l'inclusion de patients additionnels pour l'entraînement du modèle est déterminant. Nous avons donc étudié les approches faiblement supervisées. Après analyse de l'état de l'art (voir section 8.2), nous avons opté pour une méthode permettant d'apprendre à partir de quelques pixels annotés seulement dans une image. Cette approche est particulièrement intéressante pour le CaP, puisque la plupart des bases de données ont une vérité terrain basée sur la biopsie, qui ne caractérise la lésion qu'au point de biopsie et ne permet pas de connaître son étendue. De cette façon, nous avons pu inclure une base de données publique (ProstateX-2) pour laquelle la vérité terrain correspond aux coordonnées des centroïdes des lésions seulement, sans les contours précis des lésions. Nous présentons ces résultats chapitre 8.

— Les modèles que nous avons étudiés jusqu'alors sont biparamétriques, c'est-à-dire qu'ils n'incluent que le T2-w et le DWI (sous la forme des cartes ADC pour nous). Nous avons fait ce choix suite à l'étude de l'état de l'art, la quasi-totalité des modèles n'incluant pas la séquence dynamique DCE. Or, bien que cette séquence ne soit pas principale pour décider du score PI-RADS d'une lésion (voir section 2.5.2), elle contient une information qui pourrait être utile aux modèles. La principale difficulté pour l'inclusion du DCE concerne sa dimension 4D. Nous avons étudié plusieurs manières de transformer l'information initiale de 4D à 3D, en calculant différentes cartes paramétriques ou sélectionnant un volume parmi ceux acquis aux différents temps. Nous rapportons les résultats de cette étude préliminaire chapitre 9.

6.3 Chaîne de traitement des données et d'analyse des performances

6.3.1 Choix concernant la base CLARA-P

Plusieurs choix ont été faits concernant la base, ils sont décrits ci-dessous.

6.3.1.1 Patients inclus

Le premier choix a été d'exclure les patients provenant du scanner Philips Achieva. En effet, une grande partie de ces patients a été acquise avec une antenne endorectale (représentée figure 2.7B), qui induit des changements d'intensités et de contrastes par rapport aux patients acquis sans cette antenne. Lors d'une étude préliminaire, nous avons également observé que l'entraînement avec ces patients

TABLEAU 6.1 – Distribution des lésions par score de Gleason (GS) pour les 219 patients de la base CLARA-P inclus dans les expériences.

	GS 3+3	**GS 3+4**	**GS 4+3**	**GS $\geq$ 8**	**Total**
ZP	83	111	52	43	289
ZT	21	15	4	9	49
Total	104	126	56	52	338

seulement ne permettait pas d'obtenir des performances correctes, contrairement aux autres scanners. Par ailleurs, les performances des radiologues sont également plus faibles sur ce scanner (voir tableau 5.8). Après exclusion de ces 55 patients, la base contient 220 patients. En pratique, nos expériences exploitent une base de 219 patients, les contours d'un patient ayant été récupérés ultérieurement. Le tableau 6.1 présente la distribution des lésions par GS pour ces 219 patients.

6.3.1.2 Séquences incluses

Nous avons inclus les modalités T2-w et ADC en entrée des modèles, en adéquation avec la plupart des travaux de l'état de l'art (voir chapitre 4). Comme évoqué précédemment (voir section 5.2), les cartes ADC recalculées *a posteriori* ont été choisies, les cartes ADC originales étant manquantes pour certains patients qui n'auraient pas pu être inclus. Certaines études exploitent également la séquence DWI ; nous avons fait le choix de ne pas l'utiliser, les valeurs de b étant très disparates entre les patients et l'information très corrélée à celle des cartes ADC.

6.3.1.3 Choix des classes

— Les lésions GS 8 et GS $\geq$ 9 ont été regroupées dans une même classe, compte tenu du faible nombre de lésions appartenant à ces deux grades, qui correspondent à une même décision clinique.

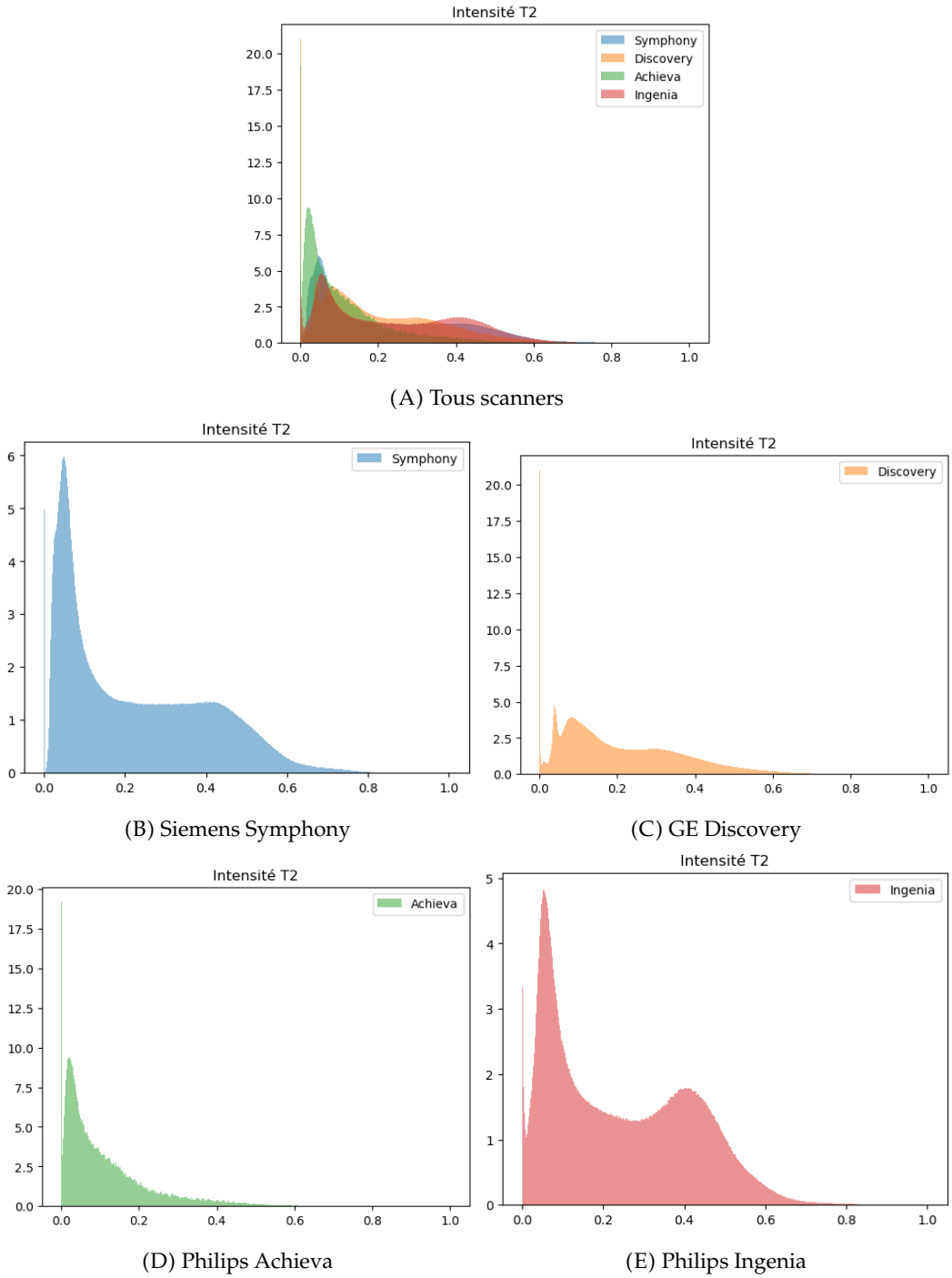
— Ayant les contours de la prostate à disposition pour tous les patients, les modèles sont entraînés à prédire la prostate en plus des différentes classes de lésions. Ainsi, davantage de pixels sont annotés. De plus, les contours de la prostate donnent un contexte spatial qui peut être utile lors de l'entraînement des modèles.

Les modèles sont donc entraînés à prédire 6 classes : la prostate, les tissus GS 6, GS 3+4, GS 4+3, GS$\geq$ 8 et le fond.

6.3.1.4 Prétraitement des images

Normalisation des intensités Étant donné la variabilité des intensités selon la machine utilisée (voir figure 5.4 pour le T2-w et figure 5.5 pour l'ADC), les intensités sont préalablement normalisées entre 0 et 1, pour les deux modalités données en entrée. Cette normalisation est d'ailleurs un prétraitement classique en analyse d'images médicales. Après normalisation, la distribution des intensités dans le T2-w est bien plus homogène entre les scanners (voir figure 6.1). Tous les scanners montrent un pic dans la distribution des intensités autour de 0.1. En revanche, un deuxième pic de plus faible fréquence autour de 0.4 est observable sur les distributions de tous les scanners sauf Philips Achieva. Les intensités de ce scanner semblent bien différentes des autres scanners, même après la normalisation dans l'intervalle

[0,1]. À noter que cette normalisation est effectuée à l'échelle du patient; un autre type de normalisation, par scanner par exemple, pourrait également être envisagé.



(A) Tous scanners

(B) Siemens Symphony

(C) GE Discovery

(D) Philips Achieva

(E) Philips Ingenia

FIGURE 6.1 – Distribution des intensités des voxels composant les volumes T2-w après normalisation des intensités par patients dans l'intervalle [0,1].

Ré-échantillonnage Comme présenté tableau 5.1, les résolutions des images diffèrent selon les séquences, les scanners et les patients. Pour assurer une cohérence spatiale, toutes les images sont sur ou sous-échantillonnées à la résolution de $1 \times 1 \times 3$ mm^3. Une meilleure résolution ne semble pas nécessaire étant donné

la taille relativement importante des lésions (voir les histogrammes figure 5.6 et le tableau 5.7). En outre, la résolution des cartes ADC est moins précise que $1 \times 1 \times 3$ mm^3, une résolution plus fine aurait nécessité une forte interpolation.

Recalage Un mouvement négligeable des patients entre les différentes séquences a été observé. Ainsi, les différentes séquences sont alignées selon les coordonnées spatiales, sans recalage supplémentaire, en accord avec les recommandations cliniques [43] et des études récentes [15, 109].

6.3.2 Base de données publiques

Afin d'évaluer la capacité de notre méthode à généraliser sur des images acquises dans un environnement clinique différent, nous avons souhaité inclure un jeu de données public externe. Cela permet également de se comparer de manière plus juste aux autres approches proposées dans la littérature. Peu de bases de données publiques pour la segmentation du CaP existent, et ce nombre se restreint davantage si l'on ne considère que les bases fournissant le GS des lésions. Par ailleurs, la plupart ne comprennent qu'un faible nombre de patients (42 pour I2CVB [78], 28 pour Prostate Fused-MRI-Pathology [89]).

Le jeu de données du challenge ProstateX-2 [82] comprend un nombre relativement important de patients (99 patients dans l'ensemble d'entraînement et 63 patients dans le jeu de test) et l'information du GS déterminé par biopsie, mais il ne fournit pas les contours des lésions comme il a été conçu pour une tâche de classification.

Cette base nous a toutefois semblé la plus adéquate pour évaluer nos modèles, d'autant plus que plusieurs travaux pour la classification, détection ou segmentation du CaP l'incluent dans leur évaluation [3, 2, 16, 27].

Base de données publique ProstateX-2

Ce jeu de données est composé d'un jeu d'entraînement de 99 patients avec 112 lésions (50 en ZP et 62 en ZT - voir tableau 6.3) et d'un jeu de test contenant 63 patients et 70 lésions. Ces données ont été acquises sur des scanners 3T Magnetom Trio et Skyra (Siemens Medical Systems) avec des paramètres d'imagerie différents (voir détails tableau 6.2), constituant ainsi deux sources différentes. La vérité terrain - obtenue par biopsie - correspond aux coordonnées du centre de chaque lésion, avec son GS associé. Les contours précis des lésions et de la prostate ne sont pas disponibles publiquement.

TABLEAU 6.2 – Paramètres d'acquisition des images incluses dans la base ProstateX-2. Le tableau 5.1 est l'équivalent pour CLARA-P.

Scanner	Champ	Séquence	T_R (ms)	T_E (ms)	FOV (mm)	Matrice (voxels)	Dimension voxel (mm)	Valeurs de b max (s/mm^2)
Siemens Skyra	3T	T2w	5660	104	192×192	384×384	$.5 \times .5 \times 3$	-
		ADC	2700	63	256×168	128×84	$2 \times 2 \times 3$	800
Siemens TrioTim	3T	T2w	4480	103	180×180	320×320	$.56 \times .56 \times 3$	-
		ADC	2700	63	212×256	106×128	$2 \times 2 \times 4$	800

TABLEAU 6.3 – Distribution des lésions par score de Gleason (GS) dans les zones périphérique (ZP) et de transition (ZT) pour les 99 patients de l'ensemble d'entraînement de la base de données ProstateX-2. Le tableau 6.1 est l'équivalent pour CLARA-P. À noter que la vérité terrain est ici la biopsie et non la pièce de prostatectomie.

	GS 3+3	**GS 3+4**	**GS 4+3**	**GS $\geq$ 8**	**Total**
ZP	14	21	9	6	50
ZT	22	20	11	9	62
Total	36	41	20	15	112

6.3.3 Stratégie d'évaluation des performances

Cette partie permet de présenter les choix effectués concernant les entraînements des modèles.

Validation croisée La base utilisée contient relativement peu de données au regard de la difficulté de la tâche, du nombre de classes à prédire (6 classes) et du faible nombre de voxels représentant les classes de lésion (voir tableau 5.7). Dans ce type de cas, les résultats peuvent être impactés par les données présentes dans les jeux d'entraînement et de validation. Il est alors conseillé de procéder par validation croisée (définie section 3.5.1), pour moyenner les résultats obtenus sur chacun des plis de validation. Dans nos expériences, la validation croisée est à 5 plis. Les patients ont été répartis dans les différents plis de manière à ce qu'ils soient le plus équilibrés possible en termes de scanner et classes de lésions représentées. Le même partitionnement a été utilisé pour toutes les expériences présentées par la suite.

Réplicats Les entraînements ont évolué au cours de la thèse. Initialement, une seule expérience de validation croisée (correspondant à une expérience sur chacun des 5 plis) était réalisée. Nous avons observé plus tard une certaine variabilité entre plusieurs validations croisées à conditions égales. Par la suite, 4 réplicats de validation croisée ont été réalisés, pour choisir le meilleur des 4 dans le chapitre 7 ou moyenner les résultats obtenus pour chacun des réplicats (chapitre 8 et 9).

Hyperparamètres Pour chaque configuration expérimentale présentée par la suite, une recherche d'hyperparamètres a été réalisée. Les hyperparamètres toujours considérés sont la vitesse d'apprentissage et la régularisation L2, optimisés conjointement. Dans le chapitre 7, le nombre de *feature maps* dans la première couche cachée ainsi que les poids de pondération des différentes classes ont également été maximisés. Les meilleures valeurs trouvées ont ensuite été conservées dans les autres expériences. Quand plusieurs termes sont présents dans la fonction de coût, leur pondération est également optimisée.

Les modèles ont été implémentés et entraînés avec les librairies Keras-TensorFlow. Dans le chapitre 7, la version 1 de TensorFlow est utilisée alors qu'il s'agit de la version 2 dans le chapitre 8 et le chapitre 9.

6.3.4 Métriques pour l'évaluation des données totalement annotées (CLARA-P)

L'évaluation des performances sur la base CLARA-P a été réalisée sur la base de métriques quantitatives de détection et de segmentation standards. Il s'agit notamment d'indices dérivés de l'analyse FROC ou des matrices de confusion, comme le score kappa de Cohen pour la tâche de détection ou l'indice de Dice pour la tâche de segmentation de la prostate. Pour le jeu de données ProstateX-2, le calcul des métriques a dû être adapté à la vérité terrain partielle.

6.3.4.1 Post-traitement des cartes de prédiction

Les modèles de segmentation étudiés dans les chapitres suivants produisent en sortie une carte à la résolution de l'image d'entrée, où chaque voxel se voit attribuer l'étiquette de classe correspondant à la valeur de probabilité maximale en sortie de la couche *softmax*. Un volume de prédiction 3D par patient est ensuite reconstruit en concaténant toutes les coupes transversales 2D de ce patient. Ceci constitue la *carte 3D brute des classes*. La *carte des lésions 3D* peut ensuite être estimée à partir de ces volumes bruts étiquetés en identifiant les composants connectés. Selon les besoins du clinicien, deux types de cartes de lésions peuvent être générés :

— les *cartes des lésions CS* (illustrées figure 6.2) sont des cartes de lésions binaires correspondant uniquement aux cancers CS. Elles sont calculées en seuillant d'abord les *cartes 3D brutes des classes* pour ne considérer que les voxels dont l'étiquette de classe correspond à un GS > 6, puis en appliquant un processus de *clustering* sur ces masques de lésions binaires CS.

— les *cartes des lésions par GS* (illustrées figure 6.3) sont des cartes de lésions multiclasses où une lésion est définie comme un groupe de voxels voisins ayant le même GS. Ceci est obtenu en appliquant le processus de *clustering* sur chaque carte par GS extraite des *cartes 3D brutes des classes*, c'est-à-dire en regroupant indépendamment tous les voxels d'une classe particulière.

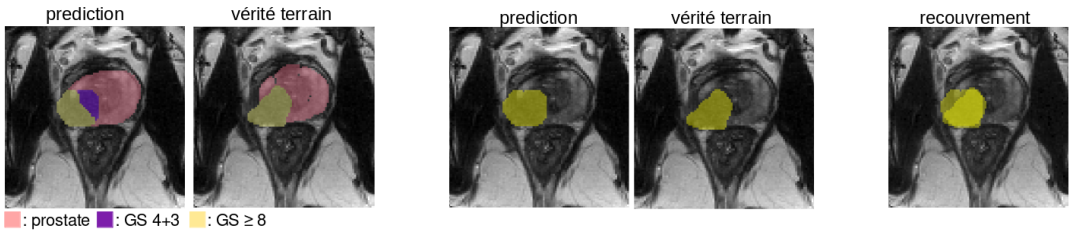


FIGURE 6.2 – Cartes de lésions CS (binaires) et recouvrement considéré pour calculer les courbes FROC.

Un post-traitement simple de ces cartes de clusters consiste à écarter les clusters plus petits qu'un certain volume prédéterminé. D'autres stratégies de post-traitement peuvent être appliquées en fonction des métriques de performance estimées sur le modèle entraîné et des besoins du clinicien. Il s'agit, par exemple, de fixer un point particulier sur les courbes FROC, correspondant à un taux moyen de faux positifs autorisé fixe et de considérer le seuil correspondant pour éliminer tous les clusters détectés ayant une probabilité lésionnelle plus faible.

Dans les chapitres suivants, nous avons considéré une règle de connectivité de 63 pour le processus de *clustering* et avons éliminé tous les clusters dont le volume est inférieur à 45 mm^3 (15 voxels) ou inférieur à 78 mm^3 (26 voxels) dans le chapitre 8.

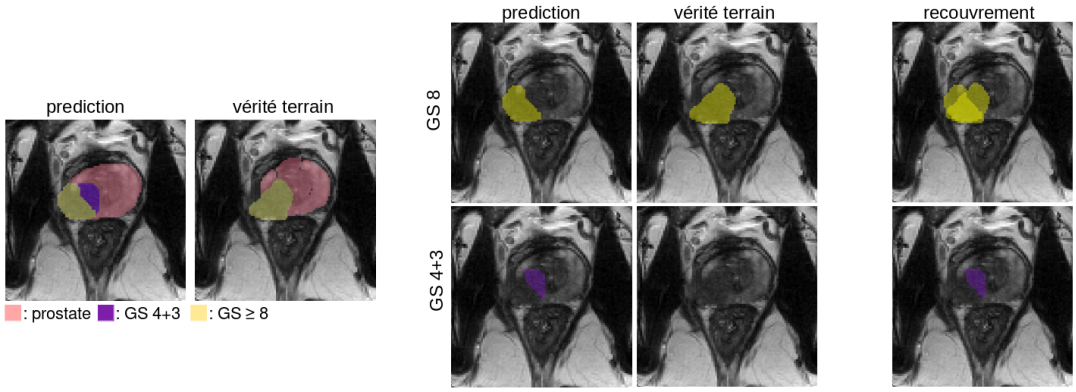


FIGURE 6.3 – Cartes de lésions GS (multiclasses) et recouvrement considéré pour calculer les courbes FROC.

Ces valeurs ont été adaptées selon le problème considéré et dérivées de la distribution de la taille des lésions de notre jeu de données privé CLARA-P, comme détaillé dans la section 5.8.

6.3.4.2 FROC pour la détection des lésions

Les performances de détection des lésions ont été évaluées par une analyse FROC basée sur les *cartes de lésions* dont le calcul est expliqué dans la section 6.3.4.1. Les courbes FROC indiquent la sensibilité de détection des lésions en fonction du nombre moyen de fausses détections par patient (voir définition section 3.5.2.3). Chaque lésion détectée s'est vu attribuer un *score de probabilité de lésion* correspondant à la moyenne des probabilités des voxels dans le cluster. La courbe FROC a ensuite été tracée en faisant varier le seuillage sur ce score de probabilité de lésion. Pour chaque valeur de ce seuil, une lésion est considérée comme un *vrai positif* (VP) lorsqu'au moins 10 % de son volume recoupe une vraie lésion et si son score de probabilité de lésion est supérieur au seuil fixé. Si elle ne recoupe pas de vraie lésion, alors la prédiction est considérée comme un *faux positif* (FP). Nous avons choisi cette valeur seuil de 10 % pour tenir compte de la variabilité inter-expert (due aux marges floues des lésions) et pour s'aligner sur d'autres études [15, 109]. À noter que grâce à la vérité terrain basée sur l'histopathologie, la définition de VP ou FP est plus précise que pour les études avec une vérité terrain basée sur la biopsie.

Deux analyses FROC différentes peuvent être réalisées à partir des *cartes de lésions* :

— la première évalue la performance des modèles pour discriminer les lésions CS (GS $>$ 6). Elle est calculée sur les *cartes de lésions CS* (voir figure 6.2), après suppression des plus petits clusters ;

— la seconde évalue la capacité du modèle à discriminer les lésions des différents GS. Elle est calculée sur les *cartes de lésions GS* (voir figure 6.3). Pour cette dernière, si une lésion a été détectée et correctement localisée mais n'a pas été assignée à la bonne classe, elle sera considérée à la fois comme un faux négatif pour sa vraie classe et comme un faux positif pour la classe prédite. Cette évaluation est donc très pessimiste.

6.3.4.3 Matrice de confusion et score kappa de Cohen pour la prédiction du GS

Pour évaluer la concordance entre la vérité terrain et la prédiction, des matrices de confusion incluant chacune des classes de GS ont été calculées.

À partir des *cartes de lésions GS*, une classe a été attribuée à chaque lésion détectée par le modèle. Si la vraie lésion chevauche plusieurs prédictions, la lésion prédite ayant le plus haut Dice avec la vérité terrain est considérée. Afin de comparer avec De Vente et al. [27], nous avons également calculé une matrice de confusion qui inclut également les FN. Dans cette configuration, les lésions qui ont été manquées par le modèle de segmentation ont été considérées comme prédites GS 6. Ensuite, le score kappa de Cohen à pondération quadratique (voir définition section 3.5.2.2) a été calculé à partir de cette matrice de confusion, comme proposé dans le challenge ProstateX-2. Pour rappel, le kappa de Cohen pondéré prend en compte la distance entre la vérité terrain et la prédiction et permet de pondérer différemment les désaccords, ce qui est utile lorsque les classes sont ordonnées. Par exemple, une lésion GS 6 prédite à tort comme GS $\geq$ 8 (désaccord élevé) sera plus pénalisée par la métrique kappa pondérée qu'une lésion GS 6 prédite à tort comme GS 3+4 (désaccord plus faible). Pour toutes les expériences, la valeur kappa rapportée est le coefficient moyen sur les 5 plis de validation.

6.3.5 Métriques pour l'évaluation des données partiellement annotées (ProstateX-2)

Les contours des lésions n'étant pas disponibles pour les données ProstateX-2, l'évaluation précédemment proposée pour les données CLARA-P n'est pas utilisable. Les métriques de segmentation et de détection ne sont pas adaptées à ce jeu de données pensé pour une tâche de classification.

Nous avons donc utilisé une métrique adaptée à une tâche de classification, qui est la matrice de confusion et le score de kappa associé. Pour adapter le calcul de cette matrice, nous avons suivi la méthode d'évaluation proposée par De Vente et al. [27]. Pour chaque lésion de la vérité terrain (dont les coordonnées sont fournies par les organisateurs du challenge ProstateX-2), elle est considérée comme un vrai positif si le centre de la lésion correspond à un voxel classé comme CS dans les cartes de lésions *CS*. La classe assignée à la lésion prédite correspond au GS le plus représenté dans le cluster correspondant. Si le centre de la lésion de référence ne recoupe aucune lésion détectée dans la *carte de lésions CS*, la lésion est rapportée comme une lésion GS 6. L'analyse des performances est ensuite réalisée sur la base des scores kappa de Cohen dérivés de la matrice de confusion avec un *bootstrap* de 1000 itérations pour plus de robustesse.

Chapitre 7

Apprentissage profond supervisé pour la segmentation du cancer de la prostate par agressivité

Ce chapitre correspond au modèle proposé avec une attention sur la ZP dans Audrey Duran, Pierre-Marc Jodoin et Carole Lartizien. « Prostate Cancer Semantic Segmentation by Gleason Score Group in bi-parametric MRI with Self Attention Model on the Peripheral Zone ». en. In : Medical Imaging with Deep Learning. *ISSN : 2640-3498. PMLR, sept. 2020, p. 193-204.*
L'extension à la prostate entière a été publiée dans Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin et Carole Lartizien. « ProstAttention-Net : A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans ». en. In : Medical Image Analysis *77 (avr. 2022), p. 102347.* ISSN *: 1361-8415.* DOI *:* 10.1016/j.media.2021.102347.

Sommaire

7.1 Introduction

Ce chapitre présente le système CADe supervisé proposé pour la détection et caractérisation du cancer de la prostate (CaP) par agressivité, entraîné et validé sur une partie de la base de données hétérogènes CLARA-P. Au début de ce travail de thèse, aucun papier n'avait été publié sur la segmentation du cancer par agressivité dans l'IRM de prostate ; la faisabilité de cette tâche restait donc à évaluer.

7.2 Méthode

7.2.1 Contexte et motivations

Le modèle proposé ici exploite les contours de la prostate disponibles, pour mettre l'attention sur la prostate. Nous avons d'abord proposé ce modèle pour la ZP seulement [35], où l'attention était focalisée sur cette zone et les lésions de la ZT ignorées. Nous l'avons ensuite étendu à la prostate entière.

Nous avons décidé de baser le modèle sur le réseau encodeur-décodeur U-Net après une analyse comparative de plusieurs réseaux de segmentation au début de la thèse. Compte tenu de ses succès en imagerie médicale, le U-Net faisait partie des modèles initialement retenus pour cette étude. Il s'est révélé être un bon compromis entre les performances et le nombre de paramètres, relativement peu élevé par rapport à d'autres modèles évalués (tel que le TiramisuNet [64]). Nous avons ensuite étudié comment améliorer les performances de ce modèle de référence et creusé la piste des modèles d'attention.

La partie suivante présente un état de l'art des mécanismes d'attention utilisés en analyse d'images médicales.

7.2.1.1 Réseaux de segmentation avec mécanismes d'attention

Les mécanismes d'attention sont couramment utilisés en vision par ordinateur pour des problèmes de classification et de segmentation [20]. Ils visent à mettre en valeur les caractéristiques discriminantes tout en atténuant celles qui sont inutiles, imitant ainsi l'attention visuelle humaine. Certains mécanismes d'attention ont montré une amélioration des performances de segmentation dans les problèmes d'imagerie médicale.

Les portes d'attention *soft* proposées dans Schlemper et al. [112] améliorent le Dice de 2-3 % par rapport au U-Net de référence pour la segmentation de 150 images abdominales 3D-CT.

Dans Roy et al. [105], l'adaptation des modules *squeeze-and-excitation* (noté SE et illustré figure 7.1) [58] pour 3 problèmes de segmentation d'images augmente le score Dice de 4 à 9 % dans le cas d'un modèle U-Net. Ces modules sont également utilisés par Rundo et al. [108] pour la segmentation des zones prostatiques sur des données IRM hétérogènes. Ils conduisent à une augmentation de 1.4 à 2.9 % du Dice

par rapport au U-Net de référence pour la segmentation de la ZP lors de l'évaluation de leur modèle multisource sur différents ensembles de données. Dans le modèle de Zhang et al. [141], la combinaison d'une attention par canal (inspirée des modules SE) et par position atteint un Dice de 1.8 % supérieur à leur référence pour la segmentation des lésions de la prostate sur des séquences IRM T2-w.

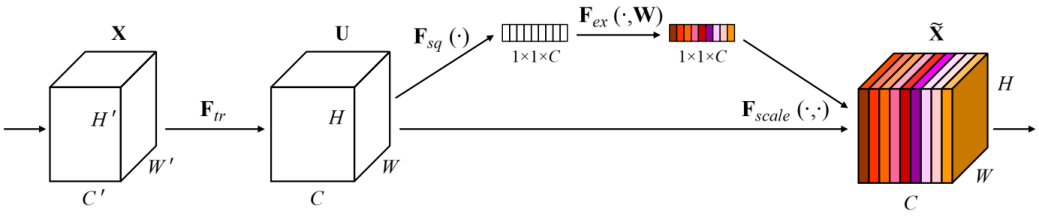


FIGURE 7.1 – Le module *squeeze-and-excitation* (noté SE) proposé dans Hu et al. [58].

De Vente et al. [27] ont testé différentes stratégies pour focaliser l'attention sur chaque zone de la prostate (ZP et ZT) séparément. Les performances de leur modèle augmentent légèrement lors de la fusion avant la convolution finale des cartes de caractéristiques zonales - correspondant à des cartes de segmentation probabiliste dérivées d'un réseau séparé - avec un score de kappa à l'échelle du voxel passant de 0.391 ± 0.062 à 0.400 ± 0.064. Toutefois, ces performances ne sont pas statistiquement meilleures que lorsque les informations zonales sont omises.

Saha et al. [109] augmentent la sensibilité à 1 FP par patient de 4.34 % par l'ajout des mécanismes SE et des portes d'attention mentionnés précédemment dans une architecture basée sur un UNet++ 3D [145] pour la segmentation binaire du CaP. Mais le gain de performance le plus important est dû à l'ajout de l'information anatomique, inclue sous forme de carte de probabilité en entrée d'un canal de leur encodeur. Cette information, qui rend compte de la prévalence spatiale et zonale des CaP CS, permet de détecter 4.10 % de lésions supplémentaires.

7.2.2 Le modèle proposé : ProstAttention-Net

Le modèle proposé - appelé ProstAttention-Net - est présenté figure 7.2. Il s'agit d'un réseau de neurones profonds entraîné de bout en bout pour réaliser simultanément deux tâches : 1) la segmentation de la prostate et 2) la détection, segmentation et classification par groupe de Gleason des lésions de la prostate.

L'encodeur du réseau extrait d'abord l'information des images T2-w et ADC présentées dans des canaux différents en entrée, jusqu'à l'espace latent. Cette représentation latente est ensuite connectée à deux branches décodeuses :

— la première va segmenter la prostate de manière binaire

— la seconde utilise la segmentation ainsi prédite pour focaliser l'attention dans la prostate et caractériser les lésions prostatiques.

La carte de probabilités obtenue en sortie du premier décodeur est utilisée comme carte d'attention *soft* pour la deuxième branche. Cette carte d'attention ainsi apprise est d'abord sous-échantillonnée à la résolution de chacun des blocs de la branche de lésion et multipliée aux cartes de caractéristiques en entrée de ce bloc. L'opération d'attention est un produit d'Hadamard effectué le long des canaux entre la sortie ré-échantillonnée de la branche prédisant la prostate et les cartes de caractéristiques de la deuxième branche. L'idée est d'aider à prédire des lésions dans la prostate en désactivant les neurones localisés en dehors de la prostate. L'attention du

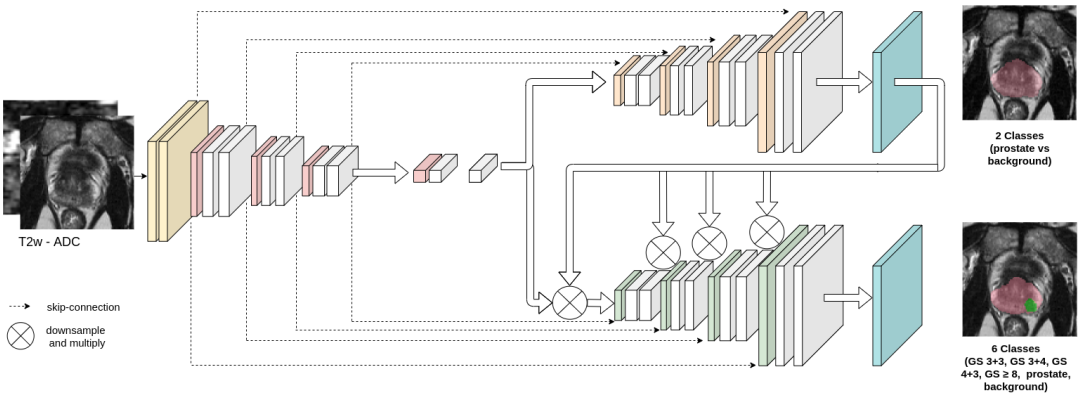


FIGURE 7.2 – ProstAttention-Net, le modèle d'attention proposé pour caractériser les lésions par GS.

réseau est donc focalisée sur la zone d'intérêt, et l'idée est également de diminuer le nombre de FP en dehors de la prostate. De plus, la segmentation de la prostate sortie par le réseau permet de visualiser les zones analysées par le réseau et peut accroître la confiance des radiologues envers le système CAD.

7.2.2.1 Architecture sous-jacente du modèle

Le modèle proposé est basé sur un U-Net [102], réseau qui a fait ses preuves pour des applications médicales mais aussi sur des images naturelles (voir section 3.4.1). La partie encodeur de notre modèle contient cinq blocs, chacun composé de deux couches convolutives avec un noyau de taille 3×3, suivies d'une couche Leaky ReLU [88] et de couches MaxPool. Nous avons inclus des couches de normalisation [59] pour réduire le sur-apprentissage. La branche du décodeur pour la prostate suit la même architecture mais avec des convolutions transposées pour augmenter la résolution des cartes de caractéristiques. Sa sortie comporte deux canaux correspondant aux classes de la prostate et du fond. Le décodeur des lésions a une architecture similaire à celle de la branche prostate, mais produit des cartes de segmentation à 6 canaux, correspondant au fond, à la prostate saine, aux lésions GS 6, GS 3+4, GS 4+3 et GS $\geq$ 8 avec des étiquettes de classe c allant de 1 à 6, respectivement.

7.2.2.2 Fonction de coût optimisée

La fonction de coût globale du ProstAttention-Net est définie comme suit :

$$L = \lambda_1 \times L_{prostate} + \lambda_2 \times L_{lesion} \tag{7.1}$$

où $L_{prostate}$ et L_{lesion} sont les fonctions de coût correspondant respectivement à la tâche de segmentation de la prostate et des lésions et λ_1, λ_2 des poids pour équilibrer les deux termes, dont la valeur peut être modifiée pendant l'apprentissage. Les fonctions de coût $L_{prostate}$ et L_{lesion} ont été définies comme la somme de l'entropie croisée et de la Dice loss, après des études préliminaires montrant les meilleures performances avec cette combinaison plutôt qu'avec l'une des deux fonctions de coût seulement. Ceci est cohérent avec d'autres travaux montrant la robustesse des fonctions de coût composées de plusieurs termes, en particulier de la Dice loss combinée à une autre fonction de coût [87]. Pour tenir compte du déséquilibre des classes, chaque terme est pondéré par un poids spécifique à la classe w_c. Ces deux fonctions

de coût peuvent être détaillées comme suit :

$$L_{prostate} = 1 - 2\frac{\sum_{c=1}^{2} w_c \sum_{i=1}^{N} y_{ci} p_{ci}}{\sum_{c=1}^{2} w_c \sum_{i=1}^{N} y_{ci} + p_{ci}} - \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} \mathbb{1}_{y_i \in C_c} w_c \log p_{ci} \tag{7.2}$$

$$L_{lesion} = 1 - 2\frac{\sum_{c=1}^{6} w_c \sum_{i=1}^{N} y_{ci} p_{ci}}{\sum_{c=1}^{6} w_c \sum_{i=1}^{N} y_{ci} + p_{ci}} - \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{6} \mathbb{1}_{y_i \in C_c} w_c \log p_{ci} \tag{7.3}$$

où w_c est le poids spécifique à la classe c, p_{ci} la probabilité prédite par le modèle pour l'observation i d'appartenir à la classe c et y_{ci} est égal à 1 si le pixel i appartient à la classe c, et 0 sinon (la vérité terrain pour le pixel i).

7.2.2.3 Détails expérimentaux

Les données d'entrée ont été prétraitées comme décrit dans la section 6.3.1.4. Chaque configuration a été évaluée en validation croisée à 5 plis, avec 4 réplicats de validation croisée pour chaque expérience compte tenu de la variabilité observée entre plusieurs expériences. Le réplicat obtenant les meilleures performances en validation est conservé. Les patients ont été répartis dans les 5 plis de manière à équilibrer autant que possible le nombre de lésions par classe et le nombre de patients de chaque scanner.

Concernant les poids de la fonction de coût définie dans l'équation 7.1, λ_1 a été fixé à 1 et λ_2 à 0 pendant les 20 premières époques afin de concentrer d'abord l'apprentissage sur la tâche de segmentation de la prostate avant de s'en servir comme attention. Ensuite, λ_2 a été fixé à 1 pour permettre une contribution égale des termes $L_{prostate}$ et L_{lesion}. Les poids spécifiques aux classes w_c ont été fixés à 0.002 pour le fond, 0.14 pour la prostate et 0.1715 pour chaque classe de lésion en fonction de la fréquence des pixels de chacune des classes estimée à partir de la base CLARA-P (voir la section 5.8).

Le réseau a été entraîné de bout en bout en utilisant Adam [72] et une régularisation L2 avec $\gamma = 10^{-4}$. La vitesse d'apprentissage initiale a été fixée à 10^{-3} avec une décroissance de 0.5 après 25 époques sans amélioration de la fonction de coût sur l'ensemble validation. Après 50 époques sans amélioration, l'entraînement est arrêté (*early stopping*). À la fin de l'entraînement, les poids associés au Dice de validation (comme dans l'équation 7.3 mais sans la classe du fond ni la pondération des classes w_c) le plus élevé sont sélectionnés. Les hyperparamètres ont été définis par recherche aléatoire sur grille. Le pipeline a été implémenté en python avec les librairies Keras-TensorFlow 1.15 [24, 1].

7.3 Analyse des performances de ProstAttention-Net

7.3.1 Expériences

ProstAttention-Net a été entraîné et testé avec la base de données de 219 patients comprenant 338 lésions décrites dans le tableau 6.1 suivant une stratégie de validation croisée (présentée section 3.5.1) à 5 plis. Nous avons répété l'expérience de validation croisée avec la meilleure combinaison d'hyperparamètres et la même répartition des données pour obtenir 4 réplicats. Comme expliqué dans le chapitre précédent, les réplicats permettent de prendre en compte la variabilité observable entre plusieurs expériences de validation croisée. L'expérience de validation croisée produisant le score kappa de Cohen (voir section 6.3.4.3) le plus élevé sur l'ensemble

de validation a été sélectionnée. Chaque pli contient 43 ou 44 patients (soit environ 1000 coupes) et est équilibré en ce qui concerne les classes de lésions et le type d'IRM présents. L'apprentissage et l'évaluation ont été effectués sur les volumes 3D entiers - chacun constitué de 24 coupes transversales - y compris des coupes sans prostate visible ou sans lésion, ce qui donne 4000 coupes d'apprentissage. À titre de comparaison, le jeu de données d'entraînement de Cao et al. [15] incluant 417 patients comprends 1400 coupes, puisque seules les coupes présentant au moins une lésion sont prises en compte. Aucun masque de prostate n'a été appliqué aux images IRM biparamétriques d'entrée ou aux cartes de prédiction de sortie, que ce soit pendant les phases d'entraînement ou d'évaluation (sauf si spécifié).

Les performances ont été évaluées d'abord de manière binaire pour rendre compte de la détection des lésions CS avec des courbes FROC, puis en multiclasse (par GS) également avec des courbes FROC mais aussi par la métrique de kappa (voir descriptions des métriques utilisées section 6.3.4). L'impact de la localisation des lésions (soit sur la ZP, soit sur la ZT) a également été évalué en appliquant les masques des zones dessinées par les radiologues pour chaque patient.

7.3.2 Résultats

7.3.2.1 Segmentation de la prostate

La segmentation de la prostate produite par la branche supérieure est évaluée en considérant le Dice moyen obtenu sur chacun des 5 jeux de validation. ProstAttention-Net obtient des résultats proches de l'état de l'art (voir section 4.2) avec un Dice de 0.875 ± 0.013.

7.3.2.2 Détection des lésions CS

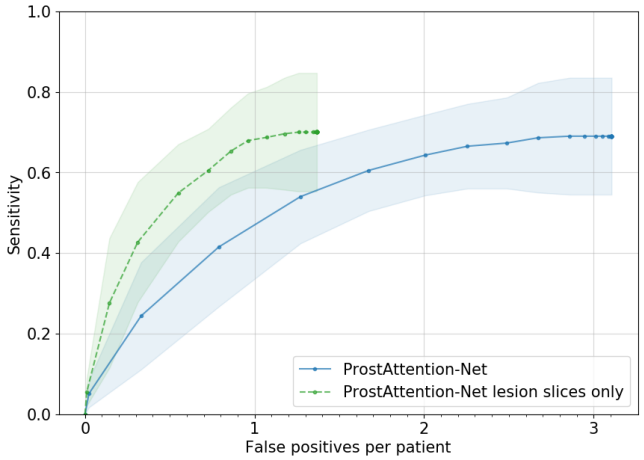
La figure 7.3 montre les courbes FROC de ProstAttention-Net pour la tâche de détection des cancers CS (GS $>$ 6). La courbe bleue montre que notre modèle atteint une sensibilité de 69.0 % $\pm$14.5 % pour 2.9 FP par patient. La courbe verte correspond aux performances de ProstAttention-Net si, au lieu de considérer l'ensemble des coupes transversales (24 coupes) du volume de chaque patient, nous n'effectuons la prédiction que sur les coupes comportant au moins une lésion, comme dans Cao et al. [15]. Cette courbe montre que notre modèle atteint une sensibilité de 68.3 % $\pm$12.1 % à 1 FP par patient. Une telle performance se compare favorablement à celle de FocalNet [15] entraîné avec l'entropie croisée. En effet, à 1 FP par patient FocalNet atteint $\sim$ 60 % avec l'entropie croisée mais $\sim$ 80 % avec la fonction de coût focale utilisée.

7.3.2.3 Segmentation par GS

Les résultats de l'analyse FROC par groupe de GS sont présentés dans le tableau 7.1. Ils reflètent la capacité du modèle à localiser les lésions mais aussi à leur attribuer le bon GS. Ils montrent que la détection des lésions des différents groupes de GS est corrélée à leur agressivité : plus le GS du cancer est élevé, meilleur est le taux de détection, comme le rapportent également Cao et al. [15]. Rappelons que ces FROC par GS sont très pessimistes car une lésion correctement localisée avec un GS incorrect est à la fois un FN pour la vraie classe et un FP pour la classe prédite.

La matrice de confusion présentée figure 7.4 montre la classe prédite pour chaque lésion détectée par rapport à la classe de référence. Nous observons que les pourcentages les plus élevés ($>$ 40 %) de lésions détectées se trouvent dans la diagonale, ce

FIGURE 7.3 – Performance du modèle ProstAttention-Net : courbes FROC évaluant la détection des lésions CS (GS $>$ 6), basées sur une validation croisée à 5 plis. La ligne bleue continue montre les performances évaluées sur le volume entier (composé de 24 coupes) et la ligne verte pointillée montre les résultats si l'on ne considère que les coupes avec au moins une lésion, comme dans Cao et al. [15]. Les zones transparentes représentent les intervalles de confiance à 95 %, correspondant à 2x l'écart type.

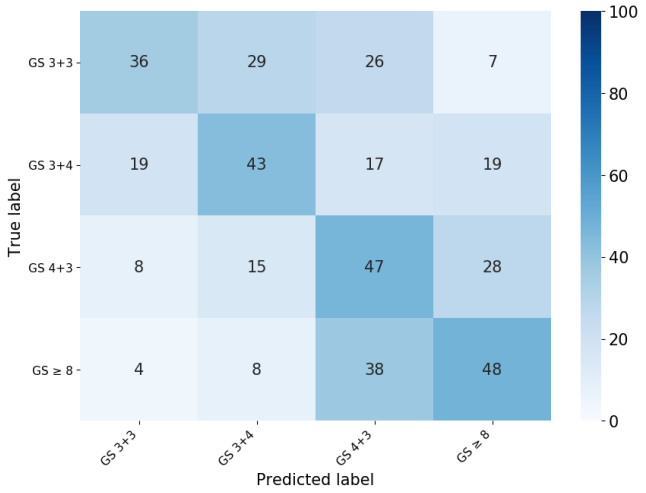


qui signifie qu'elles ont été identifiées avec le bon GS. En général, peu de lésions qui n'appartiennent pas à la classe GS $\geq$ 8 sont identifiées comme telles (7 % des GS 6, 19 % des GS 3+4 et 28 % des GS 4+3). De même, peu de lésions CS (GS $\geq$ 7) sont prédites comme des lésions GS 6 (19 % des GS 3+4, 8 % des GS 4+3 et 4 % des GS $\geq$ 8). La plupart des erreurs correspondent à des lésions GS $\geq$ 8 prédites comme GS 4+3 et à des GS 3+3 prédites comme GS 3+4 (respectivement 38 % et 29 % du taux d'erreur de classification pour ces deux classes de GS). Le score kappa de Cohen (pondéré quadratiquement) correspondant est de 0.418 ± 0.138 lorsque les résultats sont moyennés sur les 5 plis et de 0.440 lorsqu'il est calculé à partir de la matrice de confusion présentée figure 7.4.

Ces valeurs sont bonnes compte tenu de la difficulté de la tâche et se comparent favorablement aux valeurs rapportées dans l'état de l'art : elles dépassent le kappa rapporté dans De Vente et al. [27] pour une tâche de segmentation (0.172 ± 0.169 en validation croisée à 5 plis) et se rapprochent de la meilleure performance obtenue par Abraham et al. [2] pour une tâche de classification (0.4727 avec une validation croisée à 3 plis), toutes deux sur le jeu d'entraînement de ProstateX-2.

La figure 7.5 et la 7.6 montrent les cartes de prédiction extraites des *cartes brutes des classes* produites par ProstAttention-Net (ligne 4) en comparaison à la vérité terrain (ligne 3). La figure 7.5 montre des cas de réussite de notre modèle, tandis que la figure 7.6 illustre des cas d'échec. Dans le premier cas de la figure 7.5 (acquisition GE), une lésion GS 6 est présente chez le patient et visible seulement sur cette coupe. ProstAttention-Net caractérise cette lésion correctement et la délimite aussi avec précision (Dice de 0.75). Cette détection réussie est d'autant plus intéressante qu'aucun des deux radiologues n'avait identifié cette petite lésion de faible GS. Dans le second cas (acquisition Siemens), ProstAttention-Net identifie correctement la lésion de GS 3+4. La délinéation est également bonne avec un Dice de 0.62. Sur le troisième exemple (acquisition Philips), ProstAttention-Net prédit la lésion GS 4+3 avec le bon grade. On peut remarquer qu'une petite partie de la lésion est attribuée par erreur à la classe GS 3+4, ce qui est un résultat intéressant puisque les lésions GS 3+4 et

FIGURE 7.4 – Matrice de confusion normalisée de la prédiction du score de Gleason des lésions avec ProstAttention-Net. Cette matrice de confusion est la somme des 5 matrices de confusion obtenues pour chaque pli de validation. Seuls les vrais positifs (lésions détectées) sont inclus dans cette matrice. Le score kappa correspondant, pondéré par le coefficient quadratique de Cohen, est de 0.418 ± 0.138 lorsque les résultats sont moyennés sur les 5 plis de validation ou de 0.440 lorsqu'on considère cette matrice de confusion totale.



GS 4+3 sont toutes deux composées de cellules de grade 3 et de grade 4, mais avec le grade 3 ou le grade 4 respectivement majoritaire (voir définition du GS section 2.4).

Dans le premier cas de la figure 7.6 (acquisition GE), la lésion GS 3+4 est manquée alors qu'une lésion FP de GS 4+3 est prédite sur le lobe controlatéral. À noter que pour ce patient, les deux radiologues ont correctement détecté la lésion mais ont également signalé des zones suspectes supplémentaires correspondant à des FP. La deuxième colonne (acquisition Siemens) montre un patient présentant deux lésions GS 6 dans la ZP. ProstAttention-Net réussit à détecter une des deux lésions mais lui attribue le GS supérieur (GS 3+4). Ce cas de figure, où la lésion est identifiée mais avec le mauvais grade, peut néanmoins aider le radiologue à cibler la biopsie sur cette zone suspecte. Enfin, dans le dernier exemple (acquisition Philips), les deux lésions GS 3+4 et GS $\geq$ 8 situées respectivement en ZT et ZP sont manquées par le modèle alors que les radiologues n'ont pas non plus identifié la lésion GS 3+4 en ZT mais ont détecté la lésion GS $\geq$ 8 en ZP.

7.3.2.4 Performance selon la zone de la prostate considérée (ZP ou ZT)

La figure 7.7 montre les performances de ProstAttention-Net lorsque des masques anatomiques sont appliqués sur les cartes de prédiction. Comme attendu compte tenu du déséquilibre entre les lésions de la ZP et de la ZT, les performances sont plus élevées lorsque l'évaluation ne considère que la ZP plutôt que la prostate entière et atteignent une sensibilité de 70.8 ± 14.4 % à 1.5 FP par patient. Lorsqu'on ne considère que la ZT, les performances de ProstAttention-Net chutent à 56.1 ± 21.0 % de sensibilité au même nombre FP avec un écart type plus important entre les différentes expériences. Cette différence est très probablement due au très faible nombre de lésions en ZT (49) dans notre ensemble de données par rapport aux 289 lésions en ZP, comme indiqué dans le tableau 5.6. De plus, les lésions de la ZT sont difficiles à distinguer des nodules d'hyperplasie bénigne [97] et sont

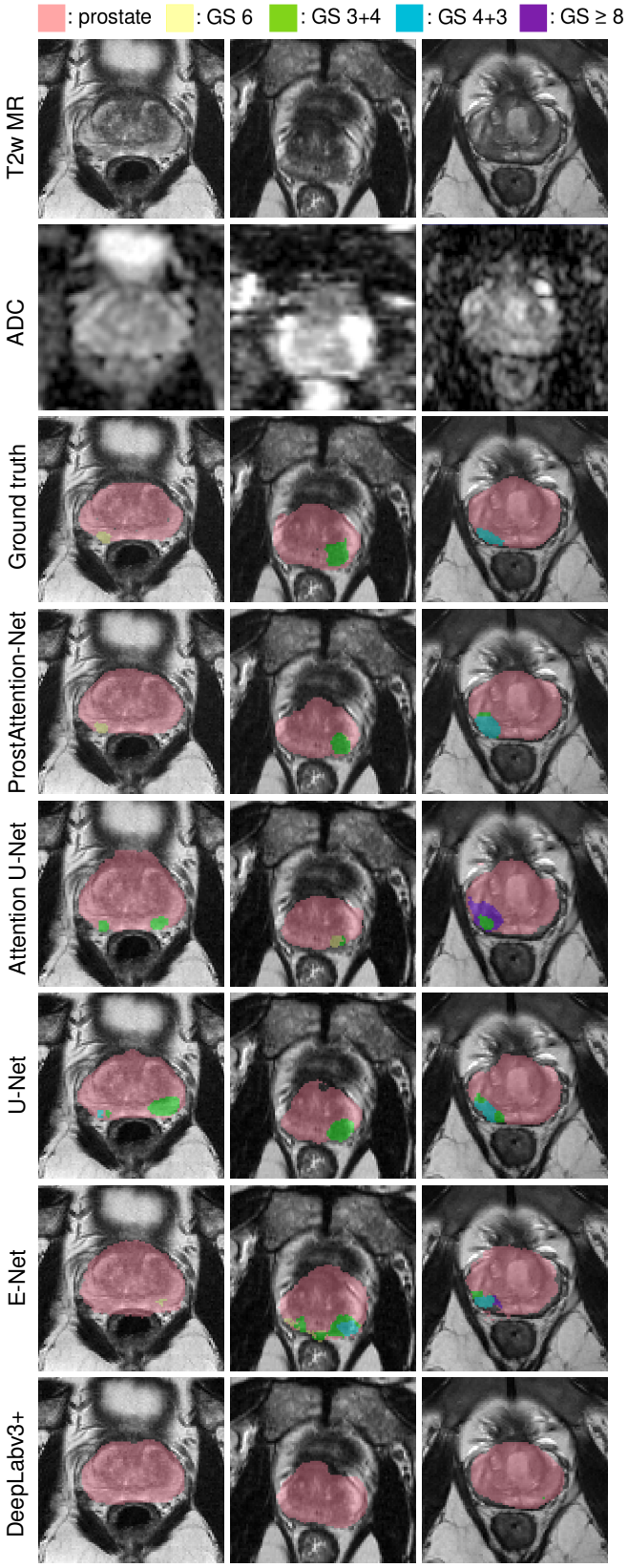


FIGURE 7.5 – Prédictions brutes pour différentes images en validation. Les images de la 1ère colonne sont issues du scanner GE 3T, celles de la 2ème colonne du Siemens 1.5T et celles de la dernière, du scanner Philips 3T. Ces exemples illustrent des cas de succès pour notre modèle (ligne 4).

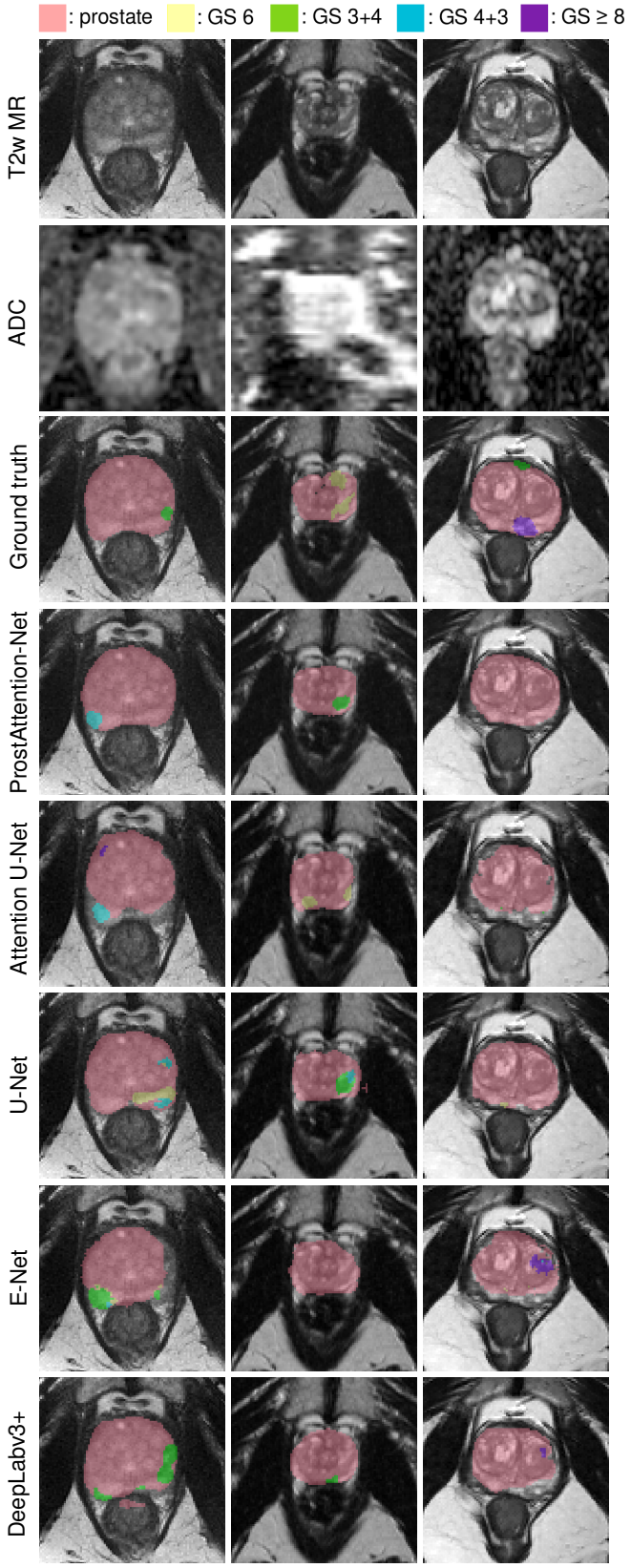
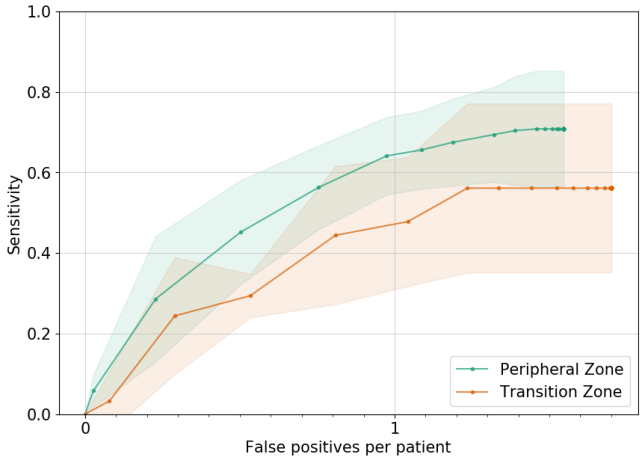


FIGURE 7.6 – Prédictions brutes pour différentes images en validation. Les images de la 1ère colonne sont issues du scanner GE 3T, celles de la 2ème colonne du Siemens 1.5T et celles de la dernière, du scanner Philips 3T. Ces exemples illustrent des cas d'échecs pour notre modèle (ligne 4).

FIGURE 7.7 – Performance selon la zone de la prostate : courbes FROC évaluant la détection des lésions CS selon la zone considérée (ZP ou ZT), basées sur une validation croisée à 5 plis. Les zones transparentes représentent les intervalles de confiance à 95 %, correspondant à 2x l'écart type.



significativement moins détectées par les radiologues que les lésions de la ZP de la base CLARA-P [10].

7.4 Comparaison à d'autres architectures issues de l'état de l'art

7.4.1 Expériences

Pour évaluer l'apport de l'architecture proposée, nous présentons les résultats de quatre autres modèles de segmentation issus de l'état de l'art :

— Le premier est le U-Net au cœur de notre architecture, sans le mécanisme d'attention.

— Le deuxième modèle est un E-Net [98], un CNN présentant un excellent compromis entre la précision et le temps d'inférence.

— Le troisième modèle est un DeepLabv3+ bâti sur un XceptionNet : [21]. DeepLabv3+ est un modèle de référence souvent utilisé par la communauté de la vision par ordinateur, qui utilise la mise en commun de pyramides spatiales et des convolutions séparables à trous pour capturer les caractéristiques multi-échelles. Nous avons également testé DeepLab-ResNet101, l'architecture utilisée dans Cao et al. [15], mais nous avons obtenu de moins bonnes performances qu'avec XceptionNet en architecture sous-jacente.

— Enfin, nous avons également inclus une comparaison avec l'Attention U-Net [112], un modèle U-Net avec des portes d'attention introduites dans les connexions résiduelles. Les portes d'attention permettent au réseau d'apprentissage de supprimer les régions non pertinentes dans une image d'entrée tout en mettant en évidence les caractéristiques saillantes utiles pour la tâche en question.

La sortie de tous les modèles comporte 6 canaux, comme la branche des lésions de ProstAttention-Net. Pour permettre une comparaison juste, toutes les

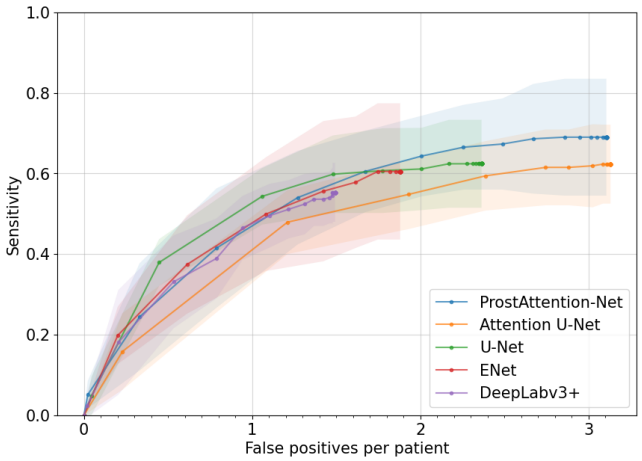
étapes du protocole expérimental ont été réalisées de la même manière que pour ProstAttention-Net, y compris le prétraitement des données, la recherche des hyper-paramètres et la sélection du meilleur modèle sur la base de 4 réplicats de validation croisée à 5 plis.

La vitesse d'apprentissage a été fixée à 10^{-3} pour U-Net et DeepLabv3+, 3.87×10^{-3} pour ENet et 1.82×10^{-2} pour Attention U-Net. En ce qui concerne le poids de régularisation, il est défini à 10^{-4} pour U-Net, DeepLabv3+ et ENet et à 1.95×10^{-5} pour Attention U-Net.

7.4.2 Résultats

La figure 7.8 compare les performances FROC des différents modèles pour la tâche de détection binaire des lésions CS. Dans l'intervalle [0, 1.5] FP, les courbes FROC se chevauchent pour la plupart et montrent des performances similaires. Ces courbes ne sont pas bornées en abscisse, de sorte qu'elles peuvent atteindre des taux de FP maximaux différents. Le modèle ProstAttention-Net atteint la sensibilité la plus élevée de 69 % pour un taux raisonnable de 3 FP par patient. ProstAttention-Net améliore également la prédiction du score de Gleason, comme le montrent les résultats de l'analyse FROC pour chaque groupe GS, rapportés dans le tableau 7.1. ProstAttention-Net atteint une sensibilité plus élevée que les autres modèles à 1 et 1.5 FP, sauf pour la classe GS 6 où ENet est plus performant (sensibilité de 24 % contre 18 % à 1.5 FP). Selon le test unilatéral de Wilcoxon, les performances de ProstAttention-Net à 1 et 1.5 FP sont toutes deux significativement supérieures à celles de U-Net et DeepLabv3+ avec une p-valeur < 0.05 pour chacun des 4 tests. Pour E-Net, la différence est juste au-dessus du seuil de significativité avec une p-valeur de 0.051 à 1 FP et 0.061 à 1.5 FP. Enfin, la différence avec Attention U-Net n'est pas non plus significative avec des p-valeurs de 0.096 et 0.058 à 1 et 1.5 FP respectivement.

FIGURE 7.8 – Comparaison des performances des différents réseaux de segmentation pour la tâche de segmentation binaire des CS CaP. Analyse FROC pour la sensibilité de détection des lésions CS (GS $>$ 6), basée sur une validation croisée à 5 plis. Les zones transparentes représentent les intervalles de confiance à 95 %, correspondant à 2x l'écart type.



Le coefficient kappa de Cohen à pondération quadratique rapporté tableau 7.2 reflète également la capacité des modèles à attribuer le bon GS. ProstAttention-Net obtient un score kappa de 0.418 ± 0.138 tandis que les scores correspondants

TABLEAU 7.1 – Sensibilité moyenne par groupe de GS à 1 et 1.5 FP par patient en validation croisée. Les différences significatives (p-valeur < 0.005) entre ProstAttention-Net et les autres modèles selon le test des rangs signés de Wilcoxon sont symbolisées par des astérisques.

	GS ≥ 8		GS 4+3		GS 3+4		GS 3+3	
	1.5 FP	1 FP	1.5 FP	1 FP	1.5 FP	1 FP	1.5 FP	1 FP
DeepLabv3+*	0.42	0.42	0.43	0.43	0.27	0.26	0.08	0.08
ENet	0.48	0.48	0.39	0.39	0.39	0.29	**0.24**	**0.15**
U-Net*	0.61	0.61	0.48	0.46	0.30	0.28	0.07	0.07
Attention U-Net	0.57	0.55	0.43	0.39	0.36	0.32	0.14	0.14
ProstAttention-Net	**0.74**	**0.70**	**0.61**	**0.52**	**0.40**	**0.36**	0.18	0.11

TABLEAU 7.2 – Score kappa de Cohen pondéré quadratiquement obtenu pour les différents modèles. Selon le test des rangs signés de Wilcoxon, le score obtenu par ProstAttention-Net n'est pas significativement supérieur à ceux des autres modèles.

DeepLabv3+	0.318 ± 0.114
ENet	0.414 ± 0.153
U-Net	0.323 ± 0.157
Attention U-Net	0.345 ± 0.158
ProstAttention-Net	$\mathbf{0.418 \pm 0.138}$

pour DeepLabv3+, U-Net et Attention U-Net sont respectivement de 0.318 ± 0.114, 0.323 ± 0.157 et 0.345 ± 0.158. Le score kappa de ENet s'avère très proche du score de ProstAttention-Net, avec une valeur moyenne de 0.414 ± 0.153. Sur la base des métriques dérivées de l'analyse FROC (tableau 7.1) et de la matrice de confusion (tableau 7.2), nous pouvons conclure que ProstAttention-Net est plus performant que les méthodes de segmentation de l'état de l'art considérées dans cette étude. Cela reflète la contribution significative du mécanisme d'attention proposé dans la segmentation des différentes classes de lésions.

Ces résultats quantitatifs sont confirmés par l'analyse qualitative de la figure 7.5 et 7.6. Globalement, les prédictions de U-Net sont les plus proches de celles de ProstAttention-Net, mais avec un échec sur le premier exemple de la figure 7.5. Sur cette même figure, Attention U-Net a pu identifier la plupart des lésions mais souvent avec un GS incorrect (premier, deuxième et troisième exemples). Cependant, c'est le seul modèle qui a pu caractériser la lésion GS 6 sur le quatrième exemple, mais en manquant toujours la deuxième lésion GS 6. Les prédictions d'E-Net semblent plus grossières (voir la large lésion prédite sur le deuxième exemple de la figure 7.5 et la mauvaise segmentation de la prostate sur les premier et troisième exemples de la figure 7.6). DeepLabv3+ est plus prudent, avec moins de lésions correctement prédites, mais aussi moins de faux positifs, comme le traduit la courbe FROC binaire (voir figure 7.8). Il est intéressant de noter que dans la première colonne de la figure 7.5, les autres modèles (en particulier U-Net et DeepLabv3+) ont obtenu de meilleurs résultats que ProstAttention-Net.

7.5 Robustesse à l'apprentissage multisource

7.5.1 Expériences

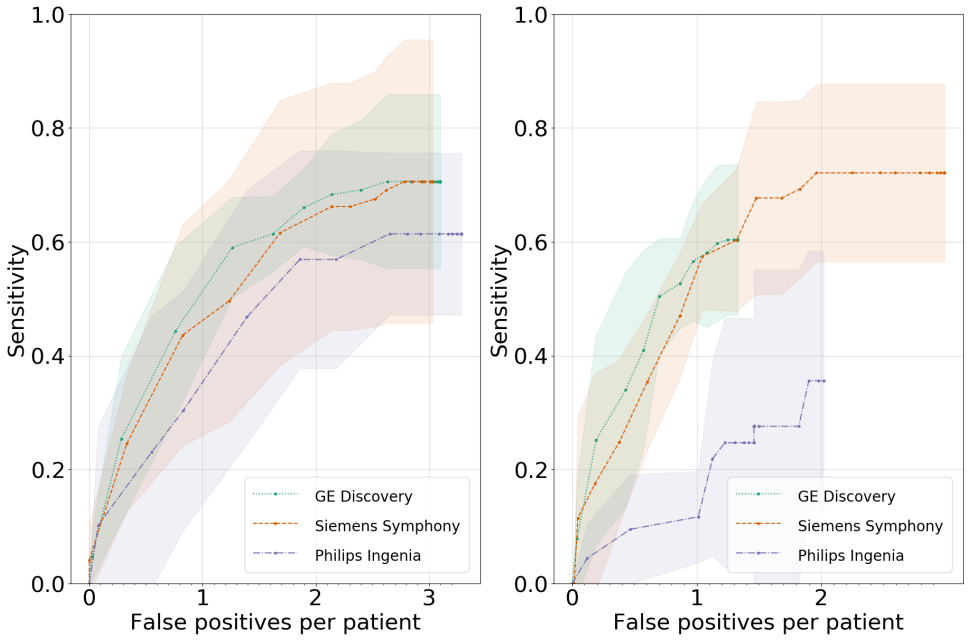
Comme indiqué dans le tableau 5.1, les données utilisées dans l'entraînement ont été acquises sur trois scanners différents, provenant de trois fabricants, avec des champs magnétiques différents et des paramètres d'imagerie disparates. De cette façon, chaque sous-ensemble de données constitue une source différente et nous pouvons nous demander si le fait d'avoir un modèle distinct pour chaque source serait plus performant qu'un modèle unique entraîné sur un ensemble de données plus grand mais hétérogène. Sur la base du modèle présenté figure 7.2, deux configurations d'entraînement ont été testées :

— **Apprentissage multisource** qui consiste à entraîner le modèle avec les 219 patients quel que soit le fabricant de l'IRM (126 GE, 67 Siemens et 26 Philips). Cette configuration correspond au modèle ProstAttention-Net entraîné selon le protocole décrit dans la section 7.3.

— **Apprentissage sur source unique** qui consiste à effectuer trois entraînements distincts de ProstAttention-Net, respectivement sur les examens GE uniquement (126 patients), les examens Siemens uniquement (67 patients) et les examens Philips uniquement (26 patients). Chacun des modèles source-spécifique a ensuite été testé sur des images acquises sur le même scanner. Chaque modèle a été entraîné et validé en utilisant une validation croisée à 5 plis. Les mêmes répartitions de patients ont été utilisées dans les expériences multisource et source unique. De cette façon, un patient affecté à un pli spécifique dans une expérience à source unique a été affecté au même pli dans les expériences à sources multiples.

7.5.2 Résultats

La figure 7.9 montre les performances du modèle pour chacun des trois scanners indépendamment. La figure 7.9A indique les performances obtenues à partir de la stratégie d'entraînement multisource, où ProstAttention-Net est entraîné avec toutes les données des trois sources correspondant à 219 examens. Les performances pour les patients GE et Siemens sont similaires ($\sim 65 - 67$ % de sensibilité à 2 FP par patient) alors que les patients Philips montrent des performances plus faibles (~ 57 % de sensibilité à 2 FP). La figure 7.9B présente les performances des trois modèles à source unique entraînés et testés sur la même source. À l'exception de Siemens, les performances chutent lorsque le modèle est entraîné sur une source unique uniquement. Cela suggère que le très petit ensemble de données Philips (26 patients) et l'ensemble de données GE (126 patients) bénéficient des données d'entraînement supplémentaires des autres scanners. La perte de sensibilité observée pour les patients Siemens dans la stratégie d'apprentissage multisource pourrait s'expliquer par le rapport de 2 :1 qui existe entre le nombre d'examens acquis sur des scanners 3T (GE Discovery et Philips Ingenia) et le scanner 1.5T (Siemens Symphony) dans le modèle multisource : le réseau pourrait en effet se concentrer un peu plus sur le modèle des données 3T, au détriment d'une altération de la sensibilité pour détecter les lésions dans les examens 1.5T Siemens.



(A) multisource (B) source unique

FIGURE 7.9 – Impact de l'apprentissage multisource. Courbes FROC pour la détection des lésions CS (GS >6) en une validation croisée à 5 plis quand les modèles sont évalués sur chacun des scanners indépendamment.
Résultats pour l'apprentissage (A) multisource (B) source unique

7.6 Comparaison à une étude préliminaire : impact de la zone d'attention et de la base de données

7.6.1 Expériences

Afin d'évaluer l'influence de la zone d'attention et de la composition de la base de données, nous avons inclus les expériences suivantes :

— **Attention sur la ZP** : ProstAttention-Net a été entraîné avec un modèle d'attention focalisé sur la zone périphérique uniquement pour se concentrer sur la détection des lésions de la ZP (correspondant à la grande majorité des lésions de la prostate).

— **Évaluation sur un jeu de données de 98 patients**, correspondant à une sous-partie de CLARA-P utilisée lorsque la base était encore incomplète (et dans une première étude, voir [35]).

7.6.2 Résultats

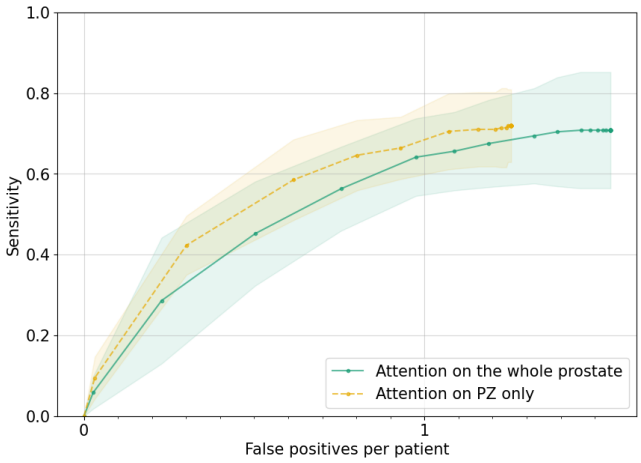
7.6.2.1 Impact de la localisation de la zone d'attention

La figure 7.10 présente l'analyse FROC pour la détection des lésions CS dans la ZP de ProstAttention-Net entraîné soit :

— avec l'attention sur la prostate entière (courbe verte) - correspondant ainsi aux résultats rapportés dans la section 7.3.2.3 ;

— avec l'attention sur la ZP uniquement (courbe jaune).

FIGURE 7.10 – Impact de la zone d'attention. Analyse FROC de la sensibilité de détection des lésions CS (GS $>$ 6) de la ZP, basée sur une validation croisée à 5 plis. ProstAttention-Net a été entraîné soit avec l'attention sur la prostate entière comme dans figure 7.3 (ligne verte pleine), soit avec l'attention sur la ZP uniquement (ligne pointillée jaune). Un masque a été appliqué sur la ZP pour omettre les lésions de la ZT dans l'analyse des performances. Les zones transparentes sont des intervalles de confiance à 95 % correspondant à 2x l'écart-type.

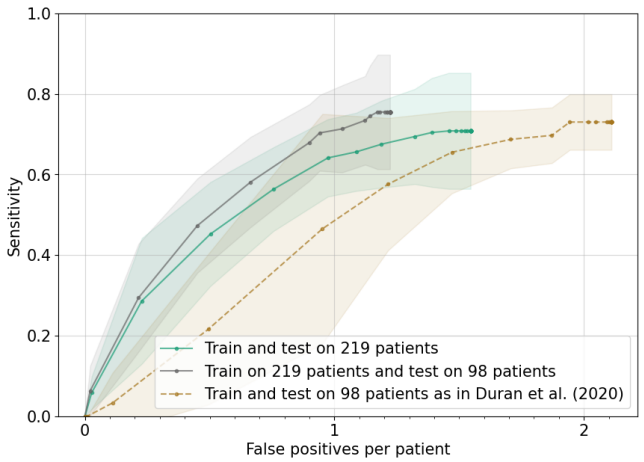


Comme dans la section 7.3.2.4, un masque a été appliqué sur la ZP de la prostate pour omettre les lésions de la ZP de l'analyse de performance. On observe que la sensibilité obtenue avec le modèle d'attention sur la ZP (courbe jaune) est supérieure à celle obtenue avec le modèle focalisant l'attention sur l'ensemble de la prostate (courbe verte) avec une sensibilité de 68.4 ± 8.4 % contre 64.6 ± 9.6 % à 1 FP. Cependant, la différence n'est pas significative et nous ne pouvons pas conclure que l'attention sur l'ensemble de la prostate dégrade les performances sur la ZP.

7.6.2.2 Impact de la base d'entraînement et de test

Nous avons testé notre modèle sur les 98 patients utilisés dans la première étude Duran et al. [35]. Pour éviter de tester le modèle sur des patients utilisés lors de la phase d'entraînement, nous avons utilisé la même répartition des patients dans les différents plis et ajouté les 121 nouveaux patients à cette base. Les résultats figure 7.11 rapportent de meilleures performances sur la base de données incluse dans Duran et al. [35], démontrant ainsi l'impact positif des données d'entraînement supplémentaires malgré une plus grande hétérogénéité de la base de 219 patients et des performances globales inférieures. Une base de données plus importante, bien que plus hétérogène, semble donc être à privilégier pour l'entraînement. La base utilisée pour tester le modèle est déterminante dans les performances rapportées, avec ici un gain d'environ 10 % lorsque le même modèle est testé sur une base réduite. Il est probable que des cas plus difficiles soient présents dans la base étendue. De plus, l'absence des Philips Ingenia dans la sous-partie de la base doit également contribuer à booster les performances, ce scanner obtenant des performances en deçà des autres (comme présenté section 7.5.2).

FIGURE 7.11 – Impact de la base de données d'entraînement et de test avec ProstAttention-Net. Analyse FROC pour la sensibilité de détection des lésions CS (GS $>$ 6), basée sur une validation croisée à 5 plis. Les lignes pleines montrent la performance de ProstAttention-Net entraîné sur les 219 patients avec une attention sur la prostate entière, testé soit sur les plis de validation provenant du même ensemble de données de 219 patients (courbe verte), soit sur les plis de validation englobant le sous-ensemble de 98 patients rapportés dans Duran et al. [35] (courbe grise). La courbe marron en pointillé indique les performances de ProstAttention-Net entraîné et testé sur le même sous-ensemble de 98 patients, l'attention étant portée uniquement sur la ZP comme dans Duran et al. [35]. Un masque a été appliqué sur la ZP pour omettre les lésions de la ZT dans l'analyse des performances. Les zones transparentes sont des intervalles de confiance à 95 % correspondant à 2x l'écart-type.



7.7 Performance sur le jeu de données public ProstateX-2

7.7.1 Expériences

Afin d'évaluer la capacité de notre méthode à généraliser sur des images acquises dans un environnement clinique différent, nous avons testé notre approche sur le jeu de données du challenge ProstateX-2 [82]. Ce jeu de données public est décrit section 6.3.2. Les données ont été acquises sur des scanners 3T Magnetom Trio et Skyra (Siemens Medical Systems) avec des paramètres d'imagerie différents (voir tableau 6.2), constituant ainsi deux sources différentes. La vérité terrain est constituée des coordonnées du centre de chaque lésion, avec son score de Gleason associé. Les contours des lésions n'étant pas disponibles pour cette base, il n'est pas possible de *fine-tuner* le modèle pré-entraîné sur ces données. Les performances de ProstAttention-Net ont été estimées sur les 99 patients du jeu d'entraînement de ProstateX-2 ; le site du challenge étant fermé, nous n'avons pas pu estimer les performances de notre réseau sur les 63 patients de test. Parmi les 5 modèles obtenus en validation croisée, celui obtenant les meilleures performances en validation sur les données CLARA-P est sélectionné pour test sur ProstateX-2. L'évaluation des performances sur ProstateX-2 a dû être adaptée à cette base de classification, comme présenté section 6.3.5.

7.7.2 Résultats

Sur l'ensemble d'entraînement de la base de données ProstateX-2, le score kappa pondéré obtenu est de 0.120 ± 0.092 (avec 1000 itérations de bootstrap). Ce kappa est inférieur à la valeur de 0.172 ± 0.169 (avec 1000 itérations de bootstrap) obtenue par De Vente et al. [27] sur les données d'entraînement de ProstateX-2, mais avec un écart type plus faible qui témoigne d'une certaine robustesse. En outre, le score de kappa rapporté par De Vente et al. [27] a été obtenu avec un modèle entraîné sur les données ProstateX-2, après qu'ils ont dessiné manuellement les masques de lésions avec un logiciel interne. Malheureusement, ces masques n'ont pas été rendus publics, ce qui ne nous permet pas de comparer équitablement ces scores. Le score de kappa obtenu avec ProstAttention-Net peut également être comparé à la valeur de kappa de 0.13 ± 0.27 rapportée par la même équipe sur l'ensemble de test ProstateX-2. Malheureusement, comme évoqué précédemment nous n'avons pas pu évaluer notre modèle sur les données de test de ProstateX-2 car le challenge est terminé.

La figure 7.12 illustre des exemples de prédictions de ProstAttention-Net sur trois images d'entraînement issues de ProstateX-2, avec des lésions en ZT. Visuellement, les résultats sur ces exemples semblent satisfaisants : la prostate est bien prédite et la segmentation des lésions de la ZT semble correcte. Sur les deuxième et troisième exemples, on peut cependant noter de petites lésions FP de GS 3+4. De plus, sur le troisième exemple, la lésion est classée à tort comme GS $\geq$ 8 au lieu de GS 4+3.

Toutefois, les performances obtenues sur le jeu de données ProstateX-2 sont inférieures à celles rapportées sur notre jeu de données privé CLARA-P. La matrice de confusion (voir figure 7.13) qui sert de base au calcul du coefficient kappa indique qu'une grande partie des lésions de ProstateX-2 détectées sont confondues avec la classe GS $\geq$ 8. Cela peut suggérer un problème d'adaptation de domaine, malgré notre modèle multisource. Les paramètres d'imagerie utilisés pour

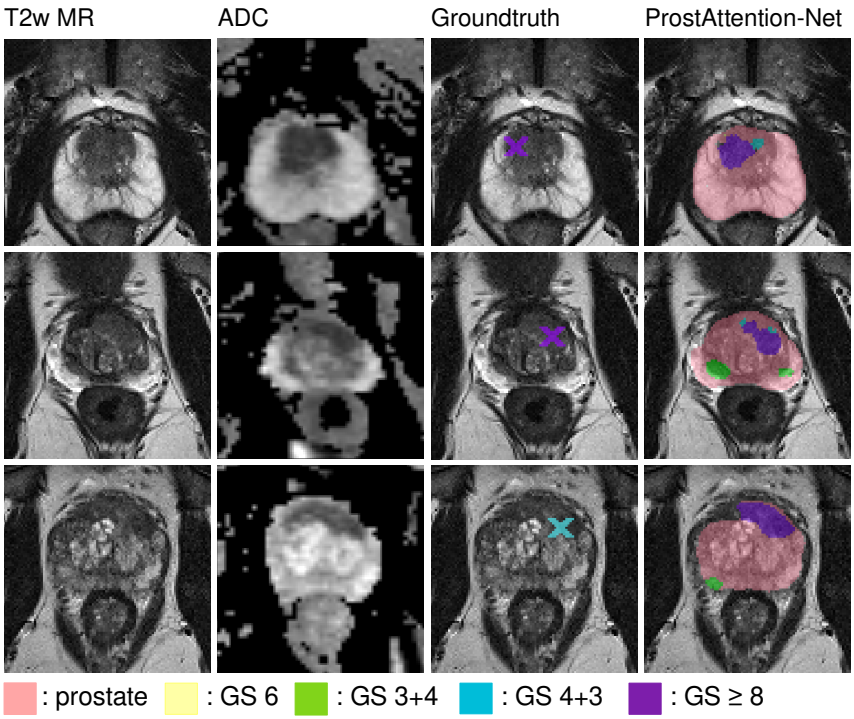


FIGURE 7.12 – Prédictions de ProstAttention-Net pour différentes images issues de la base de données publique ProstateX-2. La vérité terrain est donnée par les coordonnées du centre de la lésion, obtenues par biopsie.

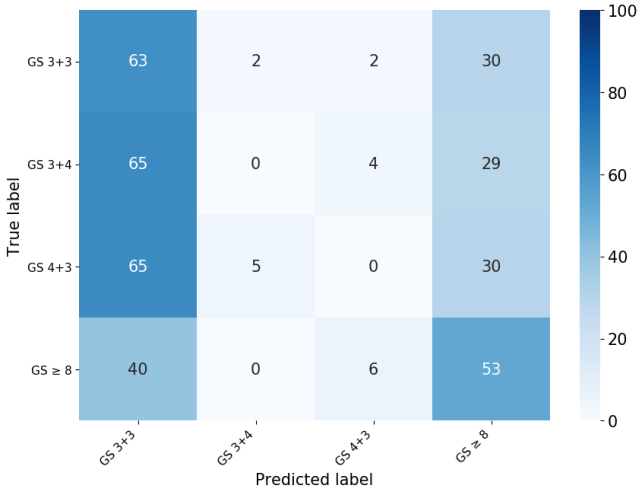


FIGURE 7.13 – Matrice de confusion normalisée de la prédiction du score de Gleason des lésions de ProstateX-2 avec le meilleur modèle de ProstAttention-Net obtenu avec la base CLARA-P. Le score kappa correspondant, obtenu après 1000 itérations de bootstrap, est de 0.120 ± 0.092.

ProstateX-2 et CLARA-P sont en effet hétérogènes et le réseau pourrait avoir besoin d'être ajusté avec les données de ProstateX-2 pour s'adapter aux caractéristiques d'intensité/texture de cet ensemble de données externe. Les lésions cancéreuses semblent généralement plus contrastées sur les cartes ADC de ProstateX-2 que sur celles de notre jeu de données. De plus, près de la moitié des lésions de ProstateX-2 se trouvent dans la ZT alors qu'elles représentent moins de 15 % des lésions (49/338) dans CLARA-P, ce qui est susceptible d'avoir un impact sur les performances puisque les performances en ZT sont inférieures à celles en ZP (comme indiqué dans la section 7.3.2.4).

7.8 Extension à un problème semi-supervisé

Le modèle proposé ci-dessus est totalement supervisé et nécessite d'avoir les contours à la fois des lésions et de la prostate pour l'entraînement, ce qui peut empêcher l'inclusion de certains patients pour lesquels une partie des annotations seulement est présente (concernant la prostate ou les lésions). Nous avons donc étudié la possibilité d'entraîner le modèle avec ce type d'annotations incomplètes.

7.8.1 Entraînement avec des données partiellement annotées

L'entraînement du modèle ProstAttention-Net peut facilement être adapté à un problème semi-supervisé, où les contours de la prostate ou des lésions seraient manquants. Ici, on considère que les contours de la prostate ou des lésions sont manquants, mais pas les deux simultanément.

Contours de la prostate manquants Les patients sans contours de la prostate disponibles peuvent être utilisés pour entraîner le modèle. Les images sont passées dans la première branche pour générer une segmentation estimée de la prostate qui sert toujours de masque d'attention *soft* dans la branche de segmentation des lésions. Ensuite, le gradient calculé grâce à la fonction de coût de la seconde branche segmentant les lésions est rétropropagé pour mettre à jour leurs poids.

Dans ce cas, la fonction de coût $L_{prostate}$ est ignorée et la fonction de coût L_{lesion} de la deuxième branche est adaptée à l'absence de la classe prostate et donc du fond également. Les voxels qui n'appartiennent à aucune des 4 classes de lésions sont alors exclus de la fonction de coût. Ceci est réalisé en faisant la somme sur les 4 classes de lésions (3 à 6, la classe 3 étant GS 6), ce qui donne la fonction de coût de la branche lésion adaptée aux annotations partielles de la prostate (APP) suivante :

$$L_{lesion-APP} = 1 - 2\frac{\sum_{c=3}^{6} w_c \sum_{i=1}^{N} y_{ci} p_{ci}}{\sum_{c=3}^{6} w_c \sum_{i=1}^{N} y_{ci} + p_{ci}} - \frac{1}{N}\sum_{i=1}^{N}\sum_{c=3}^{6} \mathbb{1}_{y_i \in C_c} w_c \log p_{ci} \tag{7.4}$$

Contours des lésions manquants Les patients dont les contours des lésions sont manquants contribueront à la fonction de coût de la branche prostatique de manière similaire. La fonction de coût de la branche lésions est adaptée pour ignorer les classes de lésions pour les patients qui n'ont pas ces contours, dans le cas d'une annotation partielle des lésions (APL). Pour ce faire, l'entropie croisée et la fonction de coût du Dice sont calculées uniquement sur les classes de non-lésion, c'est-à-dire les classes du fond ($c = 1$) et de la prostate ($c = 2$), de la même manière que dans l'équation 7.4 :

TABLEAU 7.3 – Nombre de patients annotés par scanner selon la proportion d'étiquettes disponibles.

	GE	Siemens	Total
100%	126	67	193
50%	63	33	96
30%	38	20	58
10%	12	7	19

$$L_{lesion-APL} = 1 - 2\frac{\sum_{c=1}^{2} w_c \sum_{i=1}^{N} y_{ci} p_{ci}}{\sum_{c=1}^{2} w_c \sum_{i=1}^{N} y_{ci} + p_{ci}} - \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} \mathbb{1}_{y_i \in C_c} w_c \log p_{ci} \tag{7.5}$$

Fonction de coût totale Les données d'entraînement contiennent un mélange aléatoire de coupes entièrement annotées et partiellement annotées. Ainsi, la fonction de coût finale contient quatre termes :

$$L = \lambda_1 \times L_{prostate} + \lambda_2 \times (L_{lesion} + L_{lesion-APP} + L_{lesion-APL}) \tag{7.6}$$

où les deux premiers termes concernent les coupes entièrement annotées et les deux autres les coupes partiellement annotées.

7.8.2 Matériel

Les données utilisées dans cette étude correspondent aux IRM biparamétriques de 193 patients CLARA-P imagés sur 2 scanners différents : 67 acquis sur le scanner Siemens Symphony 1.5T et 126 sur le GE Discovery 3T. Les 26 patients Philips Ingenia n'ont pas été inclus à ce moment-là dans une volonté de rester le plus proche possible de la précédente configuration de la première étude réalisée [35], où seuls des patients de GE Discovery et Siemens Symphony avaient été inclus, le reste de la base n'étant pas encore disponible.

7.8.3 Expériences

Deux séries d'expériences ont été menées pour comparer différentes stratégies d'annotations partielles :

— Le premier scénario consiste à entraîner le modèle avec des annotations de prostate manquantes pour certains patients ;

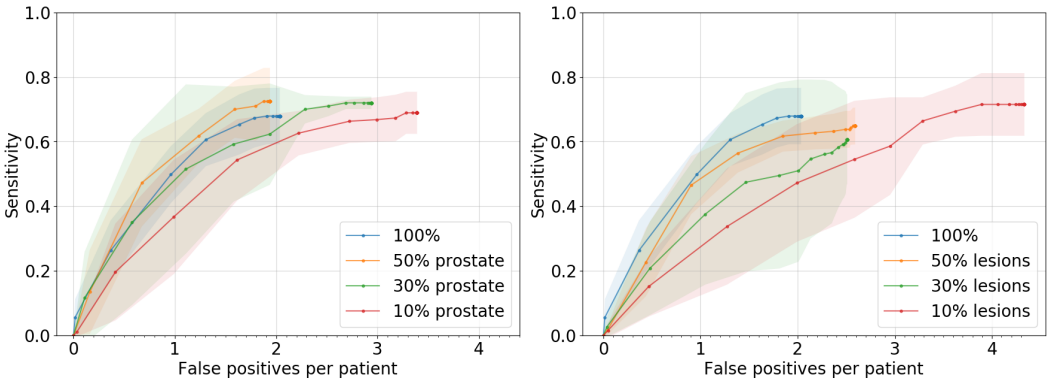
— Le second scénario consiste à entraîner le modèle avec des annotations de lésions manquantes pour certains patients. On part du postulat selon lequel l'ajout d'images d'entraînement avec des annotations de la prostate uniquement (c'est-à-dire sans annotation des lésions) pourrait améliorer les capacités de détection des lésions du système. Comme mentionné précédemment, cela découle du fait que la segmentation de la prostate est incluse dans les deux branches et génère donc un gradient dans chaque branche du réseau, branche des lésions comprise.

Pour chaque stratégie, une étude d'ablation (cf. tableau 7.3) a été réalisée pour évaluer l'effet de proportions de données partiellement annotées variables. Les patients annotés dans le jeu de données incluant 10 % de patients avec contours de la

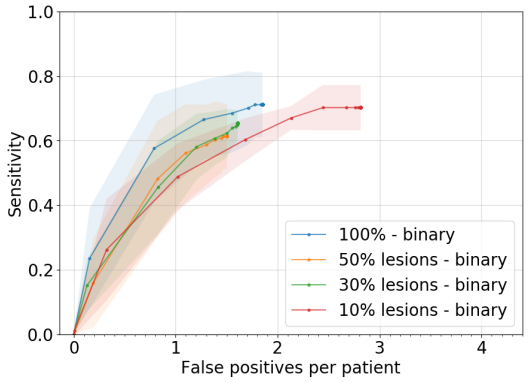
prostate sont les mêmes que dans le jeu de données incluant 10 % de patients avec les contours des lésions.

7.8.4 Résultats

La figure 7.14A témoigne de la capacité du modèle à détecter les lésions CS (GS>6) pour plusieurs pourcentages de patients ayant la prostate annotée. Comme prévu, plus le nombre de prostates annotées disponible pour l'entraînement est élevé, meilleurs sont les résultats. Cependant, on peut remarquer qu'un ensemble de données avec 50 % de prostates annotées donne d'aussi bons résultats (voire légèrement meilleurs) que l'ensemble de données annoté à 100 %. Cela suggère que le fait de disposer de 50 % du contour (c'est-à-dire 63 patients annotés pour le scanner GE et 33 pour le scanner Siemens) est suffisant pour que le modèle apprenne à segmenter la prostate et ses lésions pour les deux domaines. Avec 30 % et 10 % des annotations, les performances sont affectées avec un plus grand nombre de fausses détections et une sensibilité plus faible.



(A) Prostates partiellement annotées. (B) Lésions partiellement annotées.



(C) Lésions partiellement annotées dans un problème à 3 classes.

FIGURE 7.14 – Évaluation FROC pour la détection des lésions CS (GS>6) en validation croisée à 5 plis. Résultats pour une part variable de (A) contours de la prostate (B) contours des lésions (C) contours des lésions avec un modèle de segmentation à 3 classes.

La figure 7.14B montre les résultats des expériences pour plusieurs pourcentages de patients avec les lésions annotées. Avec 50 % des lésions annotées, la sensibilité est plus faible mais reste proche des performances obtenues avec le jeu de données entièrement annoté. À partir de 30 % de lésions annotées, on observe une baisse de performances. Le score kappa à l'échelle du voxel passe de 0.281 ± 0.069 pour le

modèle entièrement annoté à 0.209 ± 0.093, 0.186 ± 0.172 et 0.146 ± 0.191 pour les expériences à 50 %, 30 % et 10 % d'annotations respectivement. La segmentation des lésions en fonction de leur score de Gleason est une tâche difficile car les classes sont fortement déséquilibrées. Ce problème est amplifié lorsque moins de contours de lésions sont disponibles pour l'entraînement. Nous avons donc étudié l'impact du manque de contours de lésions sur un problème de segmentation à 3 classes plus facile mais plus commun, en considérant le fond, la prostate saine et les lésions CS (GS >6). Les résultats sont présentés figure 7.14C. La courbe FROC de l'expérience avec toutes les annotations se révèle proche mais au-dessus du cas à 6 classes avec une sensibilité de $68.5 \pm 12.1\%$ à 1.5 FP alors qu'elle était de 63.4 ± 9.0 % dans le cas à 6 classes. Comme pour le problème à 6 classes, les performances sont légèrement inférieures pour les modèles entraînés avec moins d'annotations : à 1.5 FP par patient, les modèles entraînés avec 50 %, 30 % et 10 % d'annotations de lésions atteignent une sensibilité de 61.3 ± 10.1 %, $62.3 \pm 7.5\%$ et ~ 57 % respectivement. La comparaison de la figure 7.14B et de la figure 7.14C montre, comme prévu, que le modèle à 3 classes est moins affecté par la suppression des annotations de lésions que le modèle à 6 classes.

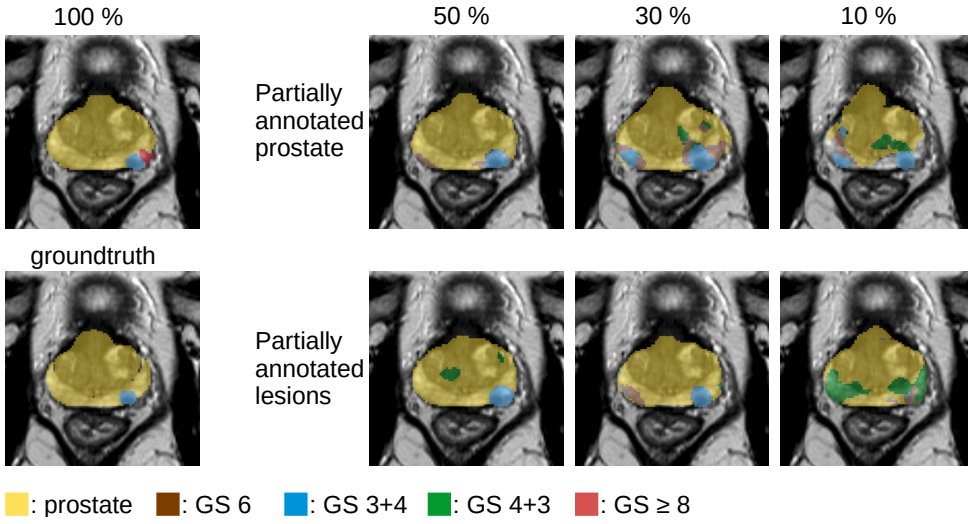


FIGURE 7.15 – Comparaison des prédictions pour les différentes expériences sur une image Siemens issue du jeu de validation.

La figure 7.15 présente des résultats visuels pour une coupe 2D sélectionnée. La première colonne montre la vérité terrain avec une lésion GS 3+4 et la prédiction obtenue par le modèle entraîné sur l'ensemble des données Siemens et GE. Les autres colonnes montrent les prédictions du même modèle avec 50 %, 30 % et 10 % des contours de la prostate (première ligne) ou des lésions (deuxième ligne). Avec le jeu de données complet, la prédiction est presque parfaite : la lésion GS 3+4 est correctement détectée (sauf une petite partie GS $\geq$ 8) et il n'y a aucun faux positif. La lésion est toujours correctement détectée avec 50 %, 30 % et même 10 % des annotations de la prostate ou des lésions mais au prix d'un nombre croissant de fausses détections (et d'une étiquette GS incorrecte pour le cas de 10 % des lésions).

7.9 Discussion

Les résultats présentés dans ce chapitre révèlent que le réseau ProstAttention-Net proposé peut segmenter avec succès la prostate (avec un Dice moyen en validation de 0.875 ± 0.013) et les lésions de la prostate avec des performances comparables à l'état de l'art. ProstAttention-Net montre également de bonnes performances dans la caractérisation du score de Gleason des lésions détectées. Nous avons montré que l'ajout de l'attention sur la prostate améliore significativement l'attribution du score de Gleason, sur la base du kappa de Cohen par lésion et des FROC par GS. Cela est cohérent avec des études récentes montrant la contribution du mécanisme d'attention et l'inclusion d'informations zonales pour la détection du CaP [109, 27]. La généralisation correcte de notre modèle sur un jeu de données externe est sûrement due en partie à l'inclusion de données multisource dans l'apprentissage. Ceci est en quelque sorte attendu et conforme à des études récentes qui rapportent également une meilleure généralisation en apprentissage multisource plutôt que source unique pour la segmentation des zones de la prostate [108, 140]. Les performances sur le jeu de données ProstateX-2 pourraient être améliorées par l'inclusion de lésions de la ZT supplémentaires dans l'apprentissage. En effet, environ 14 % seulement des lésions de la base CLARA-P sont localisées dans la ZT (49/338) alors que les lésions en ZT représentent plus de 55 % des lésions de ProstateX-2.

7.9.1 Confrontation des résultats avec l'état de l'art

À notre connaissance, seules deux études récentes ont porté sur la caractérisation du CaP par GS sur des coupes d'IRM, pour une tâche de détection [15] ou de segmentation [27] des lésions. Ces études sont présentées dans la section 4.3.4.

Comparaison à Cao et al. [15] Afin de comparer les performances rapportées dans Cao et al. [15] pour la tâche binaire de détection de lésions CS avec leur modèle FocalNet, nous avons calculé la FROC sur les coupes contenant au moins une lésion uniquement, comme cela est fait dans leur travail. Par rapport à FocalNet, la sensibilité obtenue par ProstAttention-Net à 1 FP est 10 % plus élevée que celle de FocalNet entraîné avec l'entropie croisée, mais 15 % plus faible que la sensibilité atteinte avec la combinaison des fonctions de coût focale et de *mutual finding*.

Nous avons évalué l'impact de la fonction de coût focale (notée FL) dans l'entraînement de ProstAttention-Net, en considérant deux scénarios différents :

1. Premièrement, nous avons utilisé la FL classique uniquement [81] dans la branche de décodage des lésions de ProstAttention-Net, au lieu de la combinaison de l'entropie croisée pondérée et de la Dice loss dans l'équation 7.3.
2. Deuxièmement, la FL standard a été combinée à la Dice loss pondérée.

Dans les deux cas, les performances ne se sont pas améliorées par rapport à la fonction de coût originale (données non présentées), contrairement aux résultats rapportés par Cao et al. [15]. L'une des explications possibles est que nous utilisons des versions pondérées de la CE qui tiennent déjà compte du problème de déséquilibre des classes, contrairement à l'entropie croisée (EC) standard utilisée par Cao et al. [15]. Remplacer cette EC pondérée par la FL n'a donc pas permis un gain de performance significatif. Une autre différence concerne le poids focal q décrit dans leur FL adaptée pour l'encodage ordinal, que nous n'avons pas inclus. La performance plus faible pour la détection des lésions CS pourrait également s'expliquer par le nombre

plus faible de lésions dans notre ensemble de données (rapport 2 :1). En outre, FocalNet a été entraîné et testé sur un ensemble de données plus homogène comprenant 417 patients, tous acquis sur des scanners Siemens 3T, avec les mêmes paramètres d'imagerie. Les auteurs ne rapportent pas les performances sur le jeu de données du défi ProstateX-2, ce qui rend difficile une comparaison équitable.

En ce qui concerne la comparaison pour la tâche de caractérisation du GS, la comparaison avec l'étude [15] n'est pas simple. D'après ce que nous avons compris, les mesures issues des FROC rapportées dans Cao et al. [15] ne portent que sur la détection binaire des lésions. En effet, leur FROC par GS exprime la quantité de lésions d'un GS donné détectée, sans tenir compte du GS attribué par le modèle. Cette analyse est différente de la nôtre où un vrai positif est considéré comme tel lorsqu'il recouvre une vraie lésion et qu'on lui attribue le bon score de Gleason ; sinon la lésion détectée est comptée comme un faux négatif pour cette classe et un faux positif pour la classe prédite (voir présentation des FROC section 6.3.4.2). En appliquant les mêmes règles que Cao et al. [15] pour calculer la FROC par GS, nous obtenons des sensibilités de 0.91, 0.81, 0.61 et 0.40 respectivement pour les groupes GS $\geq$ 8, GS 4+3, GS 3+4 et GS 6 (au lieu des valeurs de 0.74, 0.61, 0.40, 0.18 rapportées dans le tableau 7.1 à 1.5 FP par patient). Cao et al. [15] n'utilisent pas non plus la matrice de confusion ou le kappa de Cohen pour rendre compte des performances de prédiction du GS, ce qui rend la comparaison difficile. La capacité de leur modèle FocalNet à grader les lésions n'est évaluée que par l'intermédiaire de quatre tâches de classification binaire, présentées section 4.3.4.

De plus, notre modèle ne se contente pas de détecter des points d'intérêt comme dans Cao et al. [15] mais segmente avec précision les lésions, ainsi que la prostate entière. Nous considérons que ces deux résultats de segmentation (forme des lésions et de la prostate) sont d'un grand intérêt clinique. La segmentation de la prostate, par exemple, permet aux cliniciens de visualiser où le réseau a effectivement recherché des lésions, ce qui peut être utile pour évaluer la confiance dans le système CAD. Notre volonté de parvenir à une segmentation précise nécessite de définir des règles plus contraignantes pour définir la détection des vrais positifs, basée sur l'intersection entre la vérité terrain et la prédiction et non sur une distance entre les centroïdes, comme cela est fait dans [15]. Cela peut se faire au prix d'une légère perte de performance lors de l'évaluation des tâches de détection.

Comparaison à De Vente et al. [27] Afin de comparer avec De Vente et al. [27], nous avons calculé le score kappa de Cohen par lésion. En incluant les lésions faussement négatives dans la classe GS 6 comme ils l'ont proposé, nous avons obtenu un score kappa de 0.511 ± 0.076 pour ProstAttention-Net, ce qui est bien au-dessus de la valeur de 0.172 ± 0.169 rapportée pour leur modèle en validation croisée à 5 plis. Lorsque nous testons les capacités de généralisation de ProstAttention-Net sur l'ensemble de données d'entraînement ProstateX-2, nous obtenons un score de kappa par lésion de 0.12 ± 0.09 (avec 1000 itérations de *bootstrap*), ce qui est proche de la valeur de 0.13 ± 0.27 rapportée dans De Vente et al. [27] sur l'ensemble de test ProstateX-2 mais avec un écart-type plus faible. Rappelons que De Vente et al. [27] ont entraîné leur modèle sur l'ensemble de données d'entraînement ProstateX-2 (après en avoir contouré les lésions), de sorte que les performances rapportées sur le jeu de test ci-dessus ne correspondent pas à la performance de généralisation obtenue sur des données provenant d'un nouveau domaine, comme pour ProstAttention-Net. En outre, alors que De Vente et al. [27] ne considèrent que les lésions CS (GS $\geq$ 7), notre modèle segmente également les lésions GS 6, qui est un grade crucial à inclure en raison de la sensibilité plus faible des radiologues

pour ce grade (0.480 et 0.589 respectivement pour les deux radiologues sur les 219 patients inclus) et de l'intérêt clinique élevé pour la surveillance active.

La comparaison avec ces deux études de référence indique que les performances obtenues par ProstAttention-Net pour la segmentation et le classement GS du CaP sont similaires aux méthodes issues de l'état de l'art. Par ailleurs, notre modèle segmente la totalité des lésions (il ne s'agit pas de points comme dans Cao et al. [15]), il permet de connaître la zone de la prostate qui a été scannée par le réseau et s'avère robuste face à une base de données très hétérogène.

7.9.2 Choix des métriques

Les résultats rapportés dans cette étude ainsi que la comparaison avec l'état de l'art soulignent le fort impact de la métrique choisie pour l'évaluation des performances des modèles de détection et de segmentation multiclasses. Ce sujet est actuellement activement discuté dans la communauté [101].

Dans cette étude, nous avons mis l'accent sur la métrique du kappa de Cohen pondéré pour permettre la comparaison avec les modèles les plus avancés, mais nous avons également rapporté d'autres métriques dérivées de l'analyse FROC. La métrique kappa doit en effet être considérée avec précaution car elle est calculée à partir de la matrice de confusion qui ne contient que des détections positives. Par conséquent, un modèle qui détecte très peu de lésions (faible sensibilité) mais leur attribue le bon GS peut obtenir une valeur de kappa élevée. En outre, une classification erronée (lésion identifiée comme GS 3+4 au lieu de GS $\geq$8 par exemple) est plus pénalisée qu'une lésion manquée. Chaque lésion détectée a une plus grande influence dans la matrice de confusion que dans l'analyse FROC, ce qui peut induire le fort écart-type observé dans le score de kappa de Cohen dans une expérience de validation croisée mais aussi entre les réplicats de validation croisée.

La sensibilité par classe dérivée des courbes FROC (cf. tableau 7.1) semble plus robuste et moins sujette aux variations. Bien que cette métrique soit très sévère, nous la considérons plus significative que le coefficient kappa, car elle traduit à la fois le pourcentage de lésions identifiées et leur caractérisation. Cette métrique n'est pas non plus parfaite, car une lésion correctement détectée avec un mauvais GS sera considérée comme une erreur pour la classe de la vérité terrain et un FP pour la classe prédite.

Une métrique qui pondérerait l'erreur, comme le score kappa, mais qui prendrait également en compte la sensibilité serait de grand intérêt.

7.10 Conclusion

Dans ce chapitre, nous avons présenté un modèle d'attention multiclasse, permettant la segmentation des lésions par GS tout en exploitant les contours de la prostate disponibles. Les performances obtenues sont dans l'état de l'art et le score de kappa est le meilleur rapporté pour une tâche de segmentation. Nous avons constaté l'apport d'un modèle combinant les différentes sources plutôt qu'un modèle par source (entraîné sur moins de données en conséquence). Nous avons montré que le modèle est également adaptable à un entraînement où des contours des lésions ou de la prostate seraient manquants. Les capacités de généralisation du modèle sont à peine affectées quand seulement 50 % des contours de la prostate et 50 % des segmentations des lésions sont utilisés.

Toutefois, les performances sur le jeu de challenge externe ProstateX-2 sont nettement inférieures à celles obtenues sur notre jeu de données CLARA-P. Afin d'exploiter ces données pour lesquelles les masques ne sont pas disponibles, il serait intéressant d'exploiter des approches faiblement supervisées, permettant d'apprendre à partir de vérités terrain incomplètes.

Dans le chapitre suivant, nous évaluons un modèle faiblement supervisé pour la segmentation du CaP par GS.

Chapitre 8

Apprentissage faiblement supervisé pour la segmentation du CaP par agressivité

La seconde partie de ce chapitre (après l'état de l'art) est adaptée de Audrey Duran, Gaspard Dussert et Carole Lartizien. « Learning to segment prostate cancer by aggressiveness from scribbles in bi-parametric MRI ». In : Medical Imaging 2022 : Image Processing. *T. 12032. SPIE, 2022, p. 178-184.* DOI *:* 10.1117/12.2607502.

Sommaire

8.1 Introduction

Les modèles de segmentation supervisés impliquent d'avoir à disposition une base de données annotée à l'échelle du pixel. Or, l'annotation des données médicales est une charge difficile : tracer les contours des objets d'intérêt est un processus long et fastidieux, qui nécessite un haut niveau d'expertise médicale et une vérité terrain fiable (dans le cas du cancer de la prostate, la prostatectomie reste la référence absolue). À titre d'exemple, délimiter les deux zones de la prostate (sans les lésions) sur chacune des 24 coupes composant un volume a nécessité en moyenne 20 minutes de travail pour la base CLARA-P. Les réseaux de neurones profonds nécessitant un nombre élevé d'exemples pour apprendre correctement et généraliser, la constitution d'une base adaptée peut prendre des années.

C'est pourquoi l'apprentissage à partir de bases de données faiblement ou partiellement annotées suscite un intérêt croissant. On distingue ici l'apprentissage semi-supervisé, où une partie de la base est totalement annotée et une autre ne l'est pas du tout, de l'apprentissage faiblement supervisé, où toutes les données sont annotées mais avec une annotation dite faible ou partielle. La figure 8.1 présente les différents types d'annotations faibles possibles, réparties en 3 catégories :

— informations à l'échelle de l'image : les classes présentes dans l'image sont connues mais pas leur localisation dans l'image ;

— informations partielles : une information sur la localisation des différentes classes est présente mais de manière incomplète (boîte englobante, gribouillis, points) ;

— informations bruitées : des contours sont présents mais ils ne constituent pas une vérité fiable, des pixels n'étant pas correctement étiquetés (comme c'est le cas si des polygones servent de référence).

À noter que même si la vérité terrain est incomplète, l'objectif du réseau reste de prédire un masque de segmentation où chacun des pixels est associé à une classe. L'utilisation des fonctions de coût classiques (voir section 3.4.2), qui nécessitent de connaître la vérité terrain dans son intégralité, n'est donc pas possible.

Par la suite, nous nous concentrerons sur l'état de l'art en apprentissage faiblement supervisé dans le domaine médical.

8.2 État de l'art en imagerie médicale

Dans cette partie, nous évoquerons plusieurs travaux en imagerie médicale exploitant une vérité faible, à l'échelle de l'image ou partielle. Plusieurs revues récentes proposent une vue d'ensemble de l'état de l'art dans le domaine médical en apprentissage faible ou semi-supervisé [23, 122, 19, 99].

8.2.1 Annotations à l'échelle de l'image

L'apprentissage à partir d'annotations à l'échelle de l'image peut se séparer en deux catégories : les méthodes basées sur les cartes d'activation de classes (*class activation maps* en anglais, noté CAM) et les méthodes basées sur l'apprentissage multi-instances (*multiple instance learning*, noté ici MI).

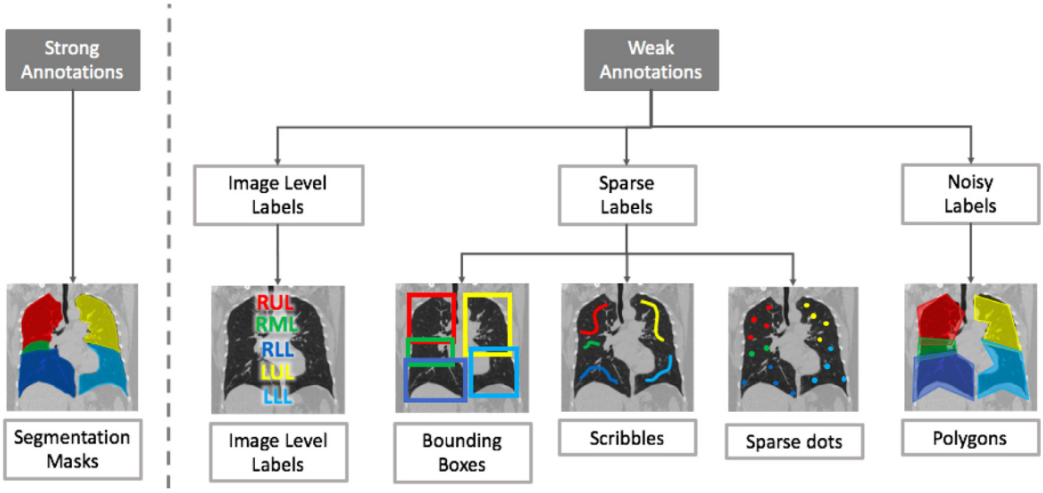


FIGURE 8.1 – Différents types d'annotations faibles possibles des lobes pulmonaires dans un scan CT de la poitrine en vue coronale. Source : Tajbakhsh et al. [122]

8.2.1.1 Les méthodes CAM

La génération de CAM a été décrite par Zhou et al. [144] et très étudiée depuis, pour les images naturelles et dans le domaine médical également. Une CAM pour une certaine catégorie indique les régions discriminantes dans l'image d'entrée utilisées par le CNN pour identifier cette catégorie. Elles permettent donc de localiser des régions d'intérêt à partir d'un modèle de classification entraîné avec des labels à l'échelle de l'image. Pour obtenir ces cartes, on ajoute habituellement un *Global Average Pooling* (noté GAP) à la fin d'un CNN ne contenant que des couches convolutives, juste avant la couche *softmax*. Ensuite, ces caractéristiques sont entrées dans une couche entièrement connectée, contenant autant de neurones qu'il y de classes dans le modèle. L'importance des régions de l'image est identifiée en rétro-projetant les poids de la couche de sortie sur les cartes de caractéristiques convolutives, une technique appelée *class activation mapping* (voir figure 8.2).

Les CAMs sont utilisées dans un contexte de localisation ou segmentation faiblement supervisée en binarisant puis ré-échantillonnant les cartes obtenues à la résolution de l'image en entrée, donnant ainsi des cartes de segmentation. Ces cartes de segmentation peuvent être ensuite utilisées comme des "pseudo vérité terrain" pour entraîner un modèle de segmentation.

Feng et al. [47] ont utilisé les CAMs précédemment décrites pour segmenter des nodules pulmonaires dans des images tomographiques en n'ayant que la classe de l'image à disposition. Un CNN de classification associé à la méthode CAM leur a permis d'obtenir des cartes des régions discriminantes, appelées *nodule activation map* (NAM). Ensuite, un CNN multi-GAP a été présenté pour profiter des NAMs provenant de couches moins profondes avec une résolution spatiale plus élevée. Leur approche a permis d'obtenir un Dice de 0.55 ± 0.33, performance se rapprochant du modèle U-Net totalement supervisé de référence, où le Dice obtenu est de 0.56 ± 0.38.

Izadyyazdanabadi et al. [62] ont proposé un modèle employant des CAMs à différentes couches/résolutions pour localiser les gliomes à partir d'images d'endomicroscopie confocale (EMC) du cerveau. Les performances du réseau sont

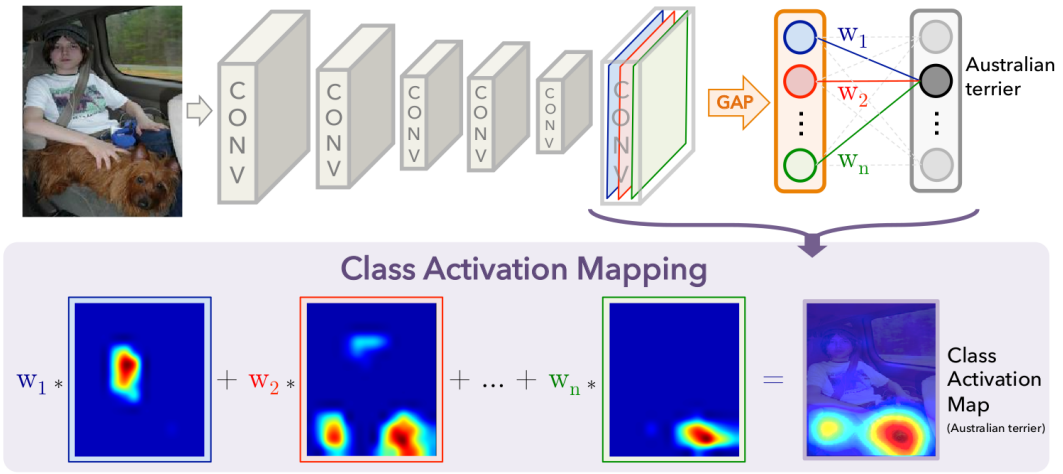


FIGURE 8.2 – *Class activation mapping* : le score de la classe prédite est renvoyé à la couche convolutive précédente pour générer les CAMs. La CAM met en évidence les régions discriminantes spécifiques à la classe. Source : Zhou et al. [144]

encore améliorées en pondérant davantage les prédictions fiables par rapport aux prédictions incertaines, où les régions activées dans la CAM d'une seule classe sont considérées fiables et celles activées dans les 2 CAMs sont dites incertaines.

Un des inconvénients des cartes de segmentation ainsi obtenues concerne leur résolution : les CAM étant calculées à la fin d'un réseau de classification, la taille de ces cartes est petite et nécessite une grande interpolation. Pour pallier ce problème, Dubost et al. [30] ont proposé un modèle appelé GP-Unet : il s'agit d'adapter le principe des CAM à un modèle de segmentation (ici une version allégée du U-Net présenté section 3.4.1, avec des connexions résiduelles), où les cartes de caractéristiques de la dernière couche ont la même résolution que l'image d'entrée. GP-Unet est utilisé pour une tâche de régression, qui consiste à prédire le nombre des petites lésions vasculaires (micro-hémorragies) dans des IRMs du cerveau. Ses performances sont comparées à 5 autres méthodes faiblement supervisées (GP-Unet sans connexions résiduelles, Gated Attention, Grad-CAM, Grad et rétropropagation guidée), à un modèle de base reposant sur l'intensité et à l'accord intra-observateur. GP-Unet rapporte la meilleure aire sous la courbe FROC pour 2 des 4 régions du cerveau (la rétropropagation guidée s'avérant meilleure pour le ganglion basal et le mésencéphale) et le moins de faux positifs pour toutes les régions. Sa sensibilité en revanche est similaire aux autres méthodes.

Les CAM ont inspiré d'autres méthodes qui en dérivent. Parmi elles, la méthode Grad-CAM [113], qui, contrairement à la méthode classique, peut être utilisée sur n'importe quel réseau et non pas seulement sur des réseaux de classification sans couche totalement connectée. La méthode Grad-CAM utilise le gradient passant dans la dernière couche de convolution d'un CNN pour évaluer l'importance de chacun des neurones dans la décision finale. La carte d'activation de classe pondérée par le gradient produit ainsi une carte thermique qui met en évidence les régions discriminantes d'une image. Guided-Grad CAM est une variante des Grad-CAM présentées dans le même travail [113], où l'on ne s'intéresse qu'aux gradients et valeurs de convolution qui sont positifs pour guider l'explication vers les centres

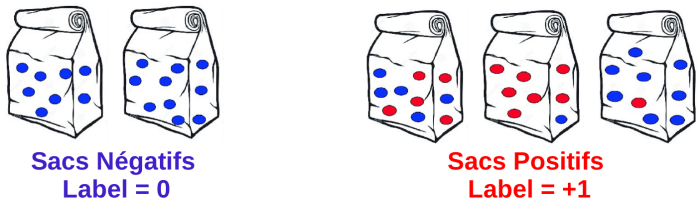


FIGURE 8.3 – Principe des méthodes MI : seule l'étiquette du sac - contenant plusieurs instances - est connue. Adaptée de Jiao et al. [67]

d'intérêt de la prédiction.

HistoSegNet, introduit dans Chan et al. [19], permet de segmenter les tissus par score de Gleason à partir de coupes histologiques en utilisant la méthode Grad-CAM. Ensuite, un post-traitement est proposé pour associer les cartes de segmentation obtenues pour chacune des classes et inclure l'arrière-plan. HistoSegNet est comparé à d'autres méthodes issues de l'état de l'art mais pour des problématiques de vision par ordinateur et non médicales. Il en ressort que seul HistoSegNet obtient de meilleures performances que la référence Grad-CAM sur les images histologiques. En revanche, HistoSegNet obtient de moins bonnes performances que les autres modèles sur des problématiques de vision par ordinateur classiques (sur des jeux de données comme PASCAL VOC2012 ou DeepGlobe).

8.2.1.2 Les méthodes MI

L'apprentissage multi-instances (MI) a été introduit par Dietterich et al. [29] pour prédire l'activité de molécules. Dans le contexte MI, les données d'apprentissage sont groupées en ensembles appelés sacs, où chaque sac contient un ensemble d'instances (voir figure 8.3). Chaque sac a une étiquette positive ou négative (dans le cas binaire), mais les instances elles-mêmes n'ont pas d'étiquette attribuée. Un sac sera négatif si toutes les instances qui le composent sont négatives et positif si au moins une instance est positive. La tâche du classifieur est ensuite de prédire le label d'un nouveau sac, ou bien des instances qui le composent. Dans tous les cas, l'apprentissage se fait à partir des données arrangées en sac. L'apprentissage MI est particulièrement utile quand il est coûteux voire impossible d'obtenir des étiquettes pour chacune des instances et répond ainsi aux problématiques de l'analyse d'images médicales.

Les méthodes MI ont été très appliquées dans le domaine médical pour l'analyse d'images histopathologiques.

Jia et al. [66] exploitent ce paradigme avec les définitions suivantes : une image histopathologique est un sac, avec comme étiquette la présence ou l'absence de cancer, et les pixels qui composent les images sont les instances. L'approche MI est utilisée sur un réseau de neurones entièrement connectés (VGG16) et appliquée à différentes couches, conjointement à une contrainte L2 sur les instances positives dont la taille doit s'approcher de l'estimation des experts. L'approche est validée sur différents jeux de données d'images histopathologiques du cancer du côlon et surpasse les quatre autres méthodes présentées, montrant l'apport de la contrainte de taille ainsi que de la supervision faible à plusieurs échelles.

Campanella et al. [13] valident l'approche MI sur une base de données conséquente comprenant 44 732 images histopathologiques (de prostate, de peau ou de métastases du sein dans les ganglions lymphatiques) provenant de 15 187 patients

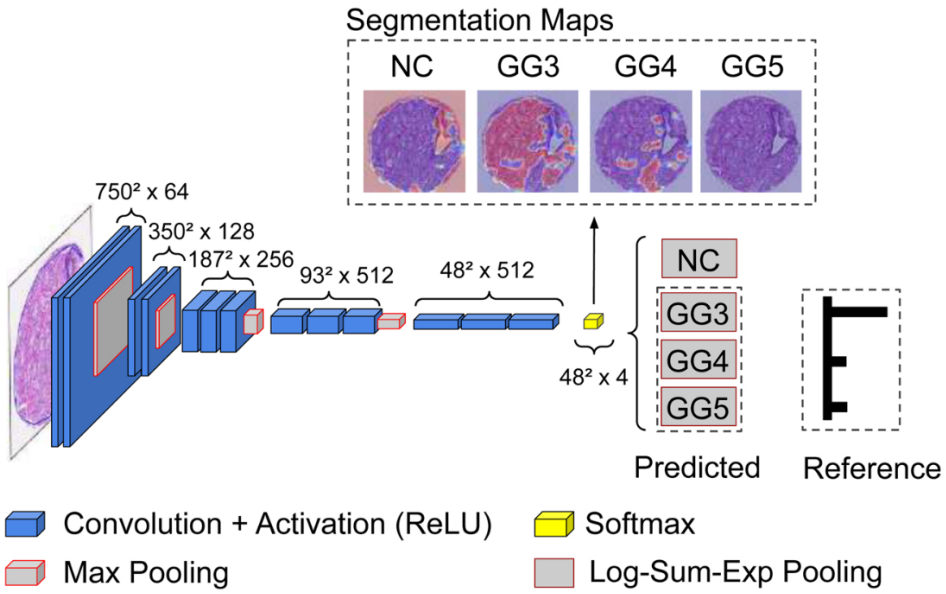


FIGURE 8.4 – WeGleNet, modèle faiblement supervisé pour la segmentation des cancers par Groupe de Gleason (noté GG) à partir d'images histopathologiques. Source : Silva-Rodríguez et al. [114]

en utilisant les rapports de diagnostic seulement. Les images ont été découpées en petites sous-parties (dites imagettes), qui sont ordonnées par un CNN selon leur probabilité d'être des instances positives. Les imagettes les plus suspectes pour chaque image sont entrées dans un réseau de neurones récurrent (RNN) pour prédire la classification finale de l'image entière. Les cartes thermiques produites par le RNN sont alors considérées comme des cartes de segmentation. Le modèle obtient une aire sous la courbe ROC supérieure à 0.98 pour tous les types de cancer évalués (prostate, sein, peau). D'après cette étude, une utilisation en clinique exclurait entre 65 et 75 % des images tout en conservant une sensibilité de 100 %.

Silva-Rodríguez et al. [114] utilisent également l'approche MI pour entraîner un CNN à segmenter les cellules par grade de Gleason avec une supervision faible, à partir d'images histologiques de prostate. De même que précédemment [66], un sac est une image histopathologique et les pixels les instances. Le modèle présenté (WeGleNet, voir figure 8.4) est composé de trois composants principaux : le classifieur - ici un VGG19 - à 4 classes (non-cancer, GG 3, GG 4, GG 5), la couche de segmentation (ou d'adaptation, similaire à une CAM) et l'opération d'agrégation. Cette dernière permet de résumer l'information présente dans toutes les cartes d'activation, comme le ferait un GAP. WeGleNet utilise à la place l'opération Log-Sum-Exp, qui obtient les meilleurs résultats.

8.2.2 Annotations partielles

Le terme d'annotations partielles fait référence à des masques incomplets, qui ne donnent la classe que d'une partie des pixels / voxels. Plusieurs méthodes permettent d'entraîner un modèle avec une telle référence, soit en générant un masque de segmentation complet pour obtenir une "pseudo vérité terrain" et entraîner un autre modèle de segmentation, soit sans compléter les masques.

L'un des premiers travaux pour l'apprentissage à partir de "gribouillis" pour l'imagerie médicale est celui de Can et al. [14]. Dans leur configuration, chacune

des classes est faiblement annotée, y compris le fond (voir figure 8.5). Avant l'entraînement du modèle, une première étape consiste à étendre les annotations à l'aide d'une méthode de marche aléatoire [53]. Ensuite, il s'agit d'optimiser les paramètres d'un CNN dont les prédictions sont ensuite combinées à un *Conditional Random Field* ou CRF (soit dense, soit un réseau de neurones récurrent). Leur modèle est validé sur deux jeux de données publics dont un jeu d'IRM de prostate, NCI-ISBI 2013. Les résultats montrent des performances proches de la supervision complète, avec un Dice de 0.722 versus 0.746 pour la ZP et 0.839 versus 0.889 pour la ZC avec leur modèle utilisant un CRF dense associé à une quantification de l'incertitude sur la segmentation.

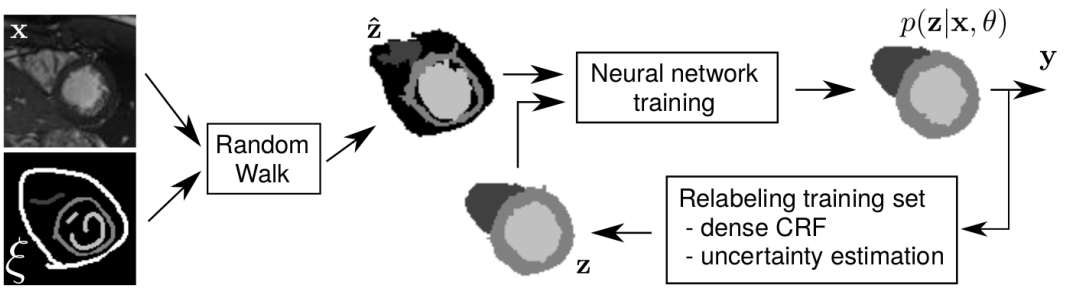


FIGURE 8.5 – Approche proposée par Can et al. [14].

Kervadec et al. [70] ont proposé une nouvelle contrainte sur la taille de la prédiction, qui pénalise une prédiction dont la taille n'est pas comprise dans un intervalle défini $[a, b]$ (voir figure 8.6). Si la classe est présente dans l'image, alors on définit $a = 1$ et $b = |\Omega|$ (le domaine image), sinon $a = b = 0$. La contrainte appliquée dépend ensuite de l'erreur quadratique entre la taille prédite et la taille minimale ou maximale attendue (voir équation 8.1).

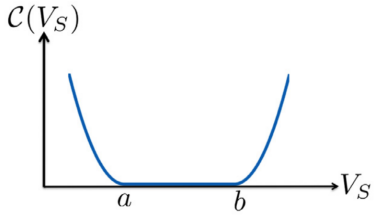


FIGURE 8.6 – Illustration de la contrainte $\mathcal{C}$ (différentiable) sur la taille de la prédiction en fonction du volume prédit V_S. $V_S = \sum_{p\in\Omega} S_p$ avec S_p la probabilité après sigmoïde prédite pour le pixel p et Ω le domaine image. Source : Kervadec et al. [70]

$$\mathcal{C}(V_S) = \begin{cases} (V_S - a)^2, & \text{si } V_S < a \\ (V_S - b)^2, & \text{si } V_S > b \\ 0, & \text{sinon} \end{cases} \tag{8.1}$$

Cette contrainte simple associée à une entropie croisée partielle (voir section 3.4.2.1) montre des résultats proches du cas totalement supervisé (qui nécessitent les masques complets) pour trois applications distinctes (segmentation de la prostate, du ventricule gauche et du corps vertébral).

Ji et al. [65] ont présenté un modèle pour segmenter différentes structures composant les tumeurs du cerveau dans des IRM, à partir d'annotation partielle. Tout d'abord, l'algorithme Graph Cut [79] permet de propager les annotations à d'autres

pixels. Un U-Net [102] est ensuite entraîné à segmenter les tumeurs à partir de cette vérité terrain partielle. La deuxième étape consiste à appliquer l'algorithme des k-moyennes pour obtenir une segmentation initiale des sous-structures composant la tumeur, qui sera utilisée pour entraîner un deuxième U-Net. Les deux réseaux sont entraînés avec l'entropie croisée partielle puis affinés avec la fonction de coût dense CRF [124]. Les résultats rapportent un Dice de 0.8823 pour la segmentation des tumeurs à l'aide des annotations initiales élargies avec Graph Cut et avec l'utilisation finale de la fonction de coût dense CRF. Cette métrique se rapproche de la référence totalement supervisée (Dice de 0.8987) et dépasse l'approche n'utilisant que les annotations partielles initiales (Dice de 0.8487).

Un autre travail récent s'est penché sur l'utilisation de points comme vérité terrain pour segmenter les noyaux cellulaires dans des images histopathologiques [137]. La classe positive est formée des pixels correspondant aux points annotés et la classe négative (le fond) comprend les pixels qui se trouvent sur les frontières du diagramme de Voronoï (obtenu en considérant les distances entre les points). Le modèle proposé (voir figure 8.7) combine un réseau entraîné à segmenter les noyaux à partir des pixels ainsi annotés et un réseau auxiliaire, PseudoEdgeNet. Ce dernier a pour but de détecter les bords des noyaux pour affiner la segmentation. Il est entraîné à produire une sortie la plus similaire possible (fonction de coût L1) à la segmentation proposée par le premier réseau, après convolution par un filtre de Sobel (connu pour la détection de contours). PseudoEdgeNet est également composé d'un module d'attention, qui va permettre de se focaliser sur le noyau. Les résultats sur la base de données publiques MoNuSeg montrent un rapport intersection sur l'union de 0.6136, supérieur à ceux obtenus avec la référence (0.5710) ou un DenseCRF (0.5813) et se rapprochant de la référence supérieure totalement supervisée (0.6522).

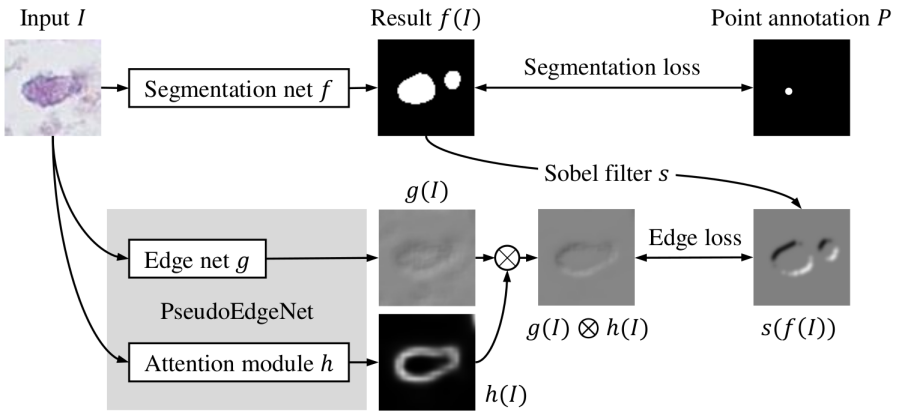


FIGURE 8.7 – Architecture proposée pour la segmentation de noyaux à partir d'une vérité faible (points). Source : Yoo et al. [137]

Pour la segmentation de cellules dans des images au microscope (toujours à partir de gribouillis), Lee et al. [76] ont proposé de générer progressivement des étiquettes fiables en combinant un pseudo-étiquetage et un filtre des étiquettes, de bout en bout. Pour cela, une moyenne exponentielle mobile est calculée pour chaque pixel de manière périodique pendant l'entraînement. Les pixels avec des prédictions suffisamment constantes (seuil variable selon les jeux de données) sont étiquetés et inclus dans la minimisation de l'entropie croisée partielle. L'évaluation sur trois jeux de données différents montre de meilleures performances avec la méthode proposée qu'avec GrabCut [80], la méthode de pseudo-étiquette proposée dans Lee et al. [75],

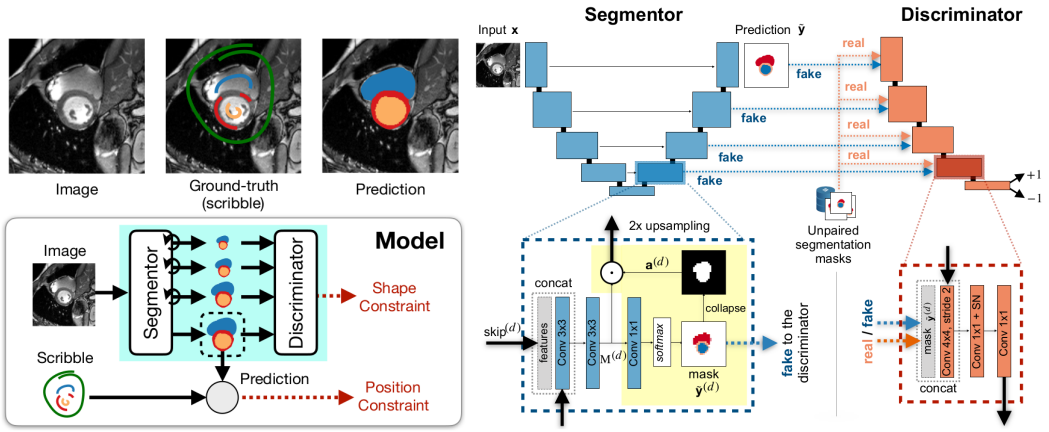


FIGURE 8.8 – (A) Vue d'ensemble et (B) architecture de la méthode proposée dans Valvano et al. [127]. Les flèches circulaires représentent le mécanisme d'attention, qui permet de supprimer les informations non pertinentes des cartes de caractéristiques. À noter que le discriminateur est entraîné à distinguer un vrai masque d'un faux à partir de masques complets, qui ne sont pas associés à l'IRM correspondant.

l'entropie croisée partielle seulement et la rLoss [124]. Toutefois, les résultats obtenus avec l'entropie partielle seulement sont déjà très corrects, avec par exemple une intersection sur l'union (IoU) de 0.9000 pour le jeu de données interne, qui passe à 0.9208 avec la méthode présentée quand la référence totalement annotée rapporte un IoU de 0.9298.

Récemment, des approches basées sur les réseaux adversaires génératifs (notés GAN) ont été proposées, toujours pour la segmentation à partir de gribouillis [142, 127]. Valvano et al. [127] ont entraîné un GAN multirésolution à générer des masques de segmentation réalistes à différentes résolutions, tout en utilisant les gribouillis pour l'apprentissage de leur localisation dans l'image (voir figure 8.8). Le réseau de segmentation utilisé est un U-Net incluant des portes d'attention antagonistes, entraîné avec une entropie croisée partielle pondérée. Le discriminateur est un encodeur convolutionnel, entraîné à différencier une prédiction du réseau de segmentation d'un vrai masque en optimisant la fonction de coût des moindres carrés adaptée aux GANs [90]. La méthode proposée nécessite donc d'avoir à disposition également des masques totalement annotés et ne repose pas uniquement sur des gribouillis. Les performances sont comparées à différentes méthodes (U-Net classique entraîné avec l'entropie croisée partielle non pondérée ou pondérée, U-Net + CRF, U-Net + post-traitement avec un auto-encodeur de débruitage, discriminateur PatchGAN [142] et U-Net entraîné avec les masques complets, avec ou sans discriminateur, la référence haute). Les résultats montrent des performances très proches de la référence totalement annotée, voire meilleures pour le jeu de données ACDC. Le Dice moyen rapporté est également meilleur que les autres méthodes comparées, sauf pour un des 4 jeux de données où la méthode de Zhang et al. [142] obtient un Dice supérieur de 0.4.

8.2.3 Application à la détection de lésions dans l'IRM de prostate

Dans cette partie, nous nous intéressons aux travaux utilisant des vérités faibles pour des tâches de détection ou segmentation du CaP à partir d'IRM.

Dans Tsehay et al. [125], la problématique consiste à segmenter les lésions cancéreuses en n'ayant à disposition que les points de biopsie. Pour générer un masque complet, tous les voxels dans un rayon de 5 mm (taille moyenne d'une lésion rapportée dans Wolters et al. [133]) autour du point de biopsie sont annotés comme cancéreux. Cette pseudo vérité terrain leur permet d'entraîner un réseau de neurones de manière totalement supervisée. Leur système CAD obtient une aire sous la courbe ROC de 0.903 ± 0.009 et prouve qu'une vérité faible peut permettre d'obtenir de bonnes performances avec un modèle supervisé, mais ne propose pas de méthode adaptée à l'apprentissage faiblement supervisé.

Wang et al. [130] ont proposé un réseau de neurones pour la détection du CaP à partir de l'annotation à l'échelle de l'image (présence / absence de cancer) seulement. Le réseau présenté est composé de deux sous-réseaux (inspirés du GoogleNet) entraînés conjointement : l'un qui extrait des caractéristiques de la séquence T2-w et l'autre des cartes ADC (voir figure 8.9). Une opération de Global Average Pooling (GAP) est appliquée à chacune des *Class Response Map* obtenues (de taille 8×8, équivalentes à des CAM) pour produire un score qui donnera la probabilité de cancer après le softmax. La fonction de coût optimisée est la somme pondérée de 3 termes : un pour la classification des images (entropie croisée), un pour la cohérence entre les prédictions des réseaux T2-w et ADC, et un dernier qui pénalise le recouvrement entre les cartes d'activation de la classe cancer et non cancer. Une évaluation par validation croisée à 5 plis sur une base de 360 patients rapporte une sensibilité de 0.6374 à 0.1 et 0.8978 à 1 FP par patients bénins. Toutefois, les CRMs restent de faible résolution (8×8 ré-échantillonnées à 299×299) et le nombre de FP rapporté assez élevé, d'autant plus qu'il n'est donné que pour les patients bénins.

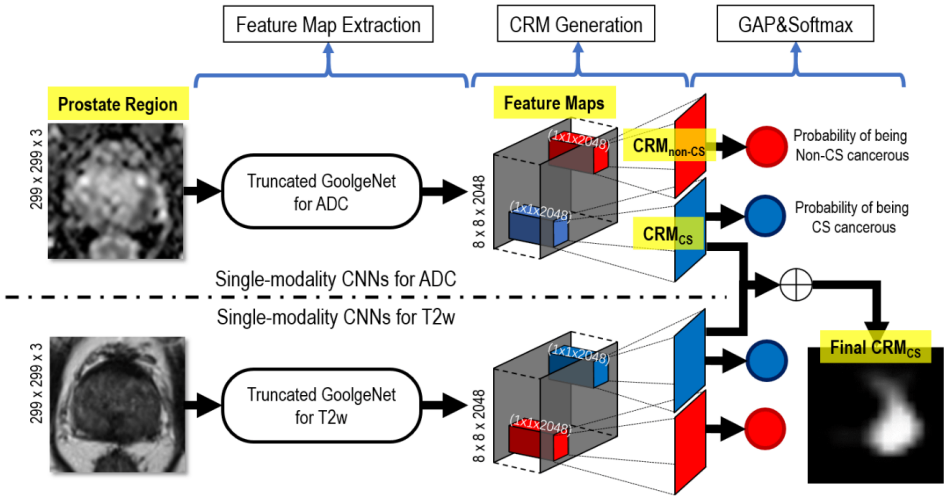


FIGURE 8.9 – Architecture du CNN multimodal présenté dans Wang et al. [130]. CRM signifie ici *Class Response Map*.

Yang et al. [136] ont récemment développé une approche spécifique à la problématique du CaP : la pièce de prostatectomie (la référence absolue) n'étant pas toujours disponible, l'idée est d'intégrer différents niveaux d'information (prostatectomie, biopsie ou rapport d'interprétation de l'IRM-mp des radiologues) dans le

modèle grâce à un réseau hiérarchique multitâche. Pour cela, le réseau est composé de trois branches :

1. une première branche associée au niveau le plus bas d'information, qui est la localisation de la lésion suspecte d'après les rapports des radiologues. Le réseau est entraîné à prédire la probabilité de cancer dans chacune des trois zones (ZP, ZT ou ZC).
2. la deuxième branche est entraînée avec la vérité issue des 12 points de biopsie (voir figure 2.5B) : chacune des 12 zones est positive si le résultat de biopsie est positif.
3. la troisième branche correspond à la vérité obtenue après l'opération de prostatectomie (voir section 2.6), qui est une référence fiable à l'échelle du voxel. Dans ce cas, la fonction de coût basée sur la métrique Dice (voir section 3.4.2.2) est optimisée.

À noter que les deux premiers niveaux d'information peuvent contenir des erreurs : le radiologue a pu repérer une zone suspecte qui ne correspond pas à une lésion dans son rapport ou en manquer une, et les points de biopsies ont également pu tomber à côté et manquer une lésion. Un lissage des étiquettes est donc appliqué pour que le réseau ne soit pas trop confiant envers cette vérité possiblement erronée.

Cette approche permet d'inclure 780 patients, parmi lesquels 121 seulement ont subi une prostatectomie, 248 ont pour référence la plus haute les résultats de biopsie et 411 patients n'ont que l'interprétation des radiologues. Les résultats montrent que l'incorporation des vérités faibles permet d'améliorer les performances par rapport à la référence avec une annotation fine seulement (AUC=0.9684 versus 0.8993 pour la référence sur la classification des patients). Toutefois, il n'est pas précisé si dans le premier cas (utilisant les 3 niveaux d'annotation), tous les patients sont inclus ou seulement les 121 patients finement annotés, pour pouvoir comparer l'apport de l'apprentissage multitâche. Il est donc difficile de conclure si le bénéfice provient de l'ajout de patients supplémentaires (pour lesquels la vérité complète n'était pas disponible) ou si les trois branches et trois niveaux d'annotation sont bénéfiques.

8.3 Méthode

Comme présenté dans la section 8.2, les approches faiblement supervisées ont été largement étudiées au cours des dernières années afin d'alléger le besoin d'annotation complète. Elles consistent à apprendre à partir d'annotations partielles, telles que l'information à l'échelle de l'image, des boîtes englobantes ou des gribouillis.

Dans cette partie, nous étudions la faisabilité de la caractérisation du CaP par agressivité à partir d'annotations partielles. Ce travail a été réalisé en collaboration avec Gaspard Dussert lors de son stage de dernière année de master cosupervisé par Carole Lartizien et moi-même. Pour rappel, la détection, segmentation et caractérisation du cancer de la prostate par agressivité n'ont été étudiées que par deux autres équipes, l'une pour la détection du cancer de la prostate [15] et l'autre pour une tâche de segmentation [27], mais toutes deux avec une approche entièrement supervisée. Outre le gain de temps avec des étiquettes faibles, ce problème est d'un grand intérêt clinique, car l'agressivité du cancer est le plus souvent déterminée à partir d'échantillons de biopsie, qui caractérisent la lésion à un endroit spécifique mais ne permettent pas de connaître l'étendue de la lésion. Ce type d'approche pourrait également permettre l'inclusion de ProstateX-2, la base de données publique contenant l'agressivité du CaP [82], mais qui ne contient que les

coordonnées du centre de la lésion (voir description de la base section 6.3.2).

Suite à l'étude de l'état de l'art en apprentissage faiblement supervisé présenté dans la section précédente, nous avons choisi de comparer :

— une approche de type Grad-CAM, qui ne nécessite que les annotations à l'échelle de l'image : celle de Dubost et al. [30] en particulier ;

— une approche utilisant des annotations partielles. Nous avons choisi celle de Kervadec et al. [70], de par la simplicité de son implémentation et les résultats concluants sur des IRM de prostate présentés dans le papier.

Les résultats préliminaires avec l'approche Grad-CAM de Dubost et al. [30] n'étant pas concluants, nous ne présentons que l'approche de Kervadec et al. [70].

8.3.1 Fonction de coût pour la supervision faible basée sur [70].

Kervadec et al. [70] ont récemment proposé une nouvelle fonction de coût pour les données partiellement annotées, présentée section 8.2.2. Elle combine un terme d'entropie croisée partielle $\mathcal{H}$ estimée sur les voxels annotés et un terme de contrainte $\mathcal{C}$, qui pénalise la prédiction dont le volume est en dehors d'un intervalle défini $[a, b]$:

$$\mathcal{H}(S) + \lambda \mathcal{C}(V_S) \tag{8.2}$$

avec λ un poids permettant d'équilibrer les deux termes de la fonction de coût. Pour rappel, la contrainte s'exprime de la manière suivante :

$$\mathcal{C}(V_S) = \begin{cases} (V_S - a)^2, & \text{si } V_S < a \\ (V_S - b)^2, & \text{si } V_S > b \\ 0, & \text{sinon} \end{cases} \tag{8.3}$$

avec a et b les bornes de l'intervalle de taille autorisé pour le volume prédit et $V_S = \sum_{p \in \Omega} S_p$ avec S_p la probabilité après sigmoïde prédite pour le pixel p et Ω le domaine image.

Nous avons adapté la fonction de coût à un problème multiclasse, avec une pondération w_c sur les classes et S_c la probabilité après softmax pour la classe c :

$$\sum_{c \in \{2,\dots,6\}} w_c \left(\mathcal{H}(S_c) + \lambda \mathcal{C}(V_{S_c}) \right) \tag{8.4}$$

Nous avons utilisé la contrainte $\mathcal{C}$ au niveau de l'image, en imposant la présence de la classe cible en fixant $a = 1$ et $b = |\Omega|$ (le domaine de l'image) ou l'absence de la cible avec les paramètres $a = b = 0$. Pour la suite, on appellera cette dernière *Tags*.

8.3.2 Architecture du modèle

Le modèle utilisé dans ce travail est basé sur un U-Net standard à quatre blocs [102], avec des couches de normalisation par *batch* pour réduire le sur-apprentissage et des activations *Leaky ReLU*. Il produit des cartes de segmentation à 6 canaux, correspondant à des étiquettes de classe c allant de 1 à 6, pour le fond, la zone globale de la prostate, les lésions GS 6, GS 3+4, GS 4+3 et GS $\geq$ 8.

8.3.3 Détails expérimentaux

Les images T2-w et ADC ont été prétraitées tel que présenté section 6.3.1.4 : l'intensité a été normalisée dans l'intervalle [0, 1], les volumes sont ré-échantillonnés à une résolution de $1 \times 1 \times 3$ mm^3 et chaque coupe est automatiquement recadrée à une taille de 96×96 pixels au centre de l'image.

Chaque configuration a été évaluée en validation croisée à 5 plis, en faisant la moyenne de 4 réplicats de validation croisée pour chaque expérience. Les patients ont été répartis dans les 5 plis de manière à équilibrer autant que possible le nombre de lésions par classe et le nombre de patients de chaque scanner et base de données (ProstateX-2 et CLARA-P) pour le cas de l'entraînement hybride. La répartition des patients de la base CLARA-P est la même que dans le chapitre 7.

Pendant la phase d'apprentissage, des techniques d'augmentation de données classiques (rotation, zoom, décalage, retournement ou *flip* horizontal) ont été appliquées afin de réduire le risque de sur-apprentissage.

Tous les réseaux ont été optimisés en utilisant Adam et une régularisation L2 avec $\gamma = 10^{-4}$. Le taux d'apprentissage initial a été fixé à 10^{-3} avec une décroissance de 0.5 après 25 époques sans amélioration de la fonction de coût en validation. Après 50 époques sans amélioration, l'entraînement est arrêté (*early stopping*). Les hyperparamètres ont été définis par recherche aléatoire sur grille. Nous avons fixé la valeur de λ pour la fonction de coût de l'équation 8.4 à 10^{-5}. Pour les modèles faiblement supervisés, les poids sur les classes sont fixés à $w_c = 0.22$ pour les classes de cancer et $w_c = 0.12$ pour la prostate ($c = 2$). Les lésions de taille inférieure à 26 voxels (c'est-à-dire 78 mm^3) ont été supprimées. Cette valeur est supérieure à la taille minimale considérée précédemment (15 voxels) car davantage de lésions sont prédites avec l'apprentissage faiblement supervisé, produisant davantage de FP.

Dans ce chapitre, la librairie Keras-TensorFlow 2.4 est utilisée.

8.4 Matériel

Deux jeux de données différents sont utilisés dans cette étude, contenant tous deux des informations sur l'agressivité de la lésion (par score de Gleason, noté GS) :

— Base CLARA-P : 219 examens IRM de patients avec l'annotation fine au niveau du pixel obtenus grâce à la pièce de prostatectomie radicale (voir description de la base chapitre 5) ;

— Base de challenge ProstateX-2 : 99 examens IRM de patients avec les coordonnées du centre de chaque lésion basées sur les résultats de la biopsie comme vérité terrain (voir description section 6.3.2).

Les séquences T2-w et ADC sont considérées dans cette étude.

Des annotations faibles correspondant à des disques ont été générées automatiquement pour les bases CLARA-P et ProstateX-2 :

— CLARA-P : disques de centre aléatoire de rayon ≤ 4 pixels (c'est-à-dire 4 mm) qui correspondent à chaque lésion et à la prostate. La taille du rayon est diminuée s'il est difficile de faire rentrer le disque dans la lésion.

— ProstateX-2 : disques de rayon = 4 pixels (soit 4 mm) aux coordonnées du centre de la lésion, fournies par le challenge. Comme les annotations de la prostate ne sont pas disponibles, elles ont été déduites de la position de la lésion : la coordonnée x du centre du disque de la prostate a été choisie à 11

TABLEAU 8.1 – Ratio des voxels annotés pour les différentes classes des données CLARA-P.

Classe	**Prostate**	**GS 6**	**GS 3+4**	**GS 4+3**	**GS $\geq$ 8**	**Total**
Ratio (%)	4.92	32.77	29.36	25.85	18.36	6.35

mm du centre de la lésion en direction du centre de l'image. Pour la position y, elle a été définie à 11 mm en direction du centre de l'image si la lésion se trouvait dans la ZT et est restée inchangée si la lésion se trouvait dans la ZP. Les annotations dessinées dans la coupe contenant le centre de la lésion ont été reportées sur les deux coupes adjacentes, et seules les coupes annotées ont été utilisées pour l'entraînement. Dans les cas (peu fréquents) où plusieurs lésions sont présentes sur une coupe ou des coupes adjacentes, la prostate est annotée autant de fois qu'il y a de lésions. La cohérence des annotations obtenues a été vérifiée visuellement.

8.5 Évaluation de la fonction de coût faiblement supervisée

8.5.1 Expériences

La première expérience consiste à valider l'approche basée sur la fonction de coût de Kervadec et al. [70] pour la segmentation de lésions par agressivité. Pour cela, elle est comparée à l'approche entièrement supervisée sur les 219 patients de la base totalement annotée CLARA-P. L'apport de la contrainte par rapport à l'entropie croisée partielle est également évalué. Dans l'approche supervisée, le modèle est entraîné avec la somme de l'entropie croisée et de la Dice loss pondérées (voir l'équation 7.3). Une fois validée, cette approche nous permet d'inclure les données faiblement annotées du challenge ProstateX-2 dans l'entraînement du modèle.

8.5.2 Résultats

Méthode	Jeu de données d'entraînement	Performance sur CLARA-P (val)			Performance sur Px2
		Kappa	Sensi à 2FP	Dice prostate	Kappa
Totalement supervisé	CLARA-P	$\mathbf{0.324 \pm 0.053}$	$\mathbf{0.649 \pm 0.033}$	0.799 ± 0.004	0.013 ± 0.082
EC partielle	CLARA-P	-0.054 ± 0.193	0.016 ± 0.008	0.081 ± 0.006	-0.005 ± 0.019
EC partielle + Tags	CLARA-P	0.289 ± 0.072	0.618 ± 0.044	0.800 ± 0.017	0.047 ± 0.060
EC partielle + Tags	Px2	0.134 ± 0.144	0.026 ± 0.012	0.121 ± 0.010	-0.002 ± 0.003
EC partielle + Tags	CLARA-P + Px2	0.262 ± 0.061	0.587 ± 0.053	$\mathbf{0.802 \pm 0.005}$	$\mathbf{0.276 \pm 0.037}$
Cao et al. [15]	privé	-	~ 0.89	-	-
De Vente et al. [27]	Px2	-	-	-	0.172 ± 0.169

TABLEAU 8.2 – Performances de segmentation et comparaison avec l'état de l'art. Nos résultats (quatre premières lignes) correspondent à la moyenne des métriques obtenues sur 4 réplicats de la validation croisée 5 fois. EC : entropie croisée. Px2 : ProstateX-2. Tags : Contrainte basée sur la présence ou l'absence de l'objet dans l'image. À noter que dans [15], seules les coupes contenant au moins une lésion sont incluses dans l'analyse des performances, de sorte que la sensibilité à un taux donné de faux positifs par patient n'est pas comparable à notre calcul de la sensibilité estimée sur la base de toutes les coupes des patients (24 coupes en moyenne).

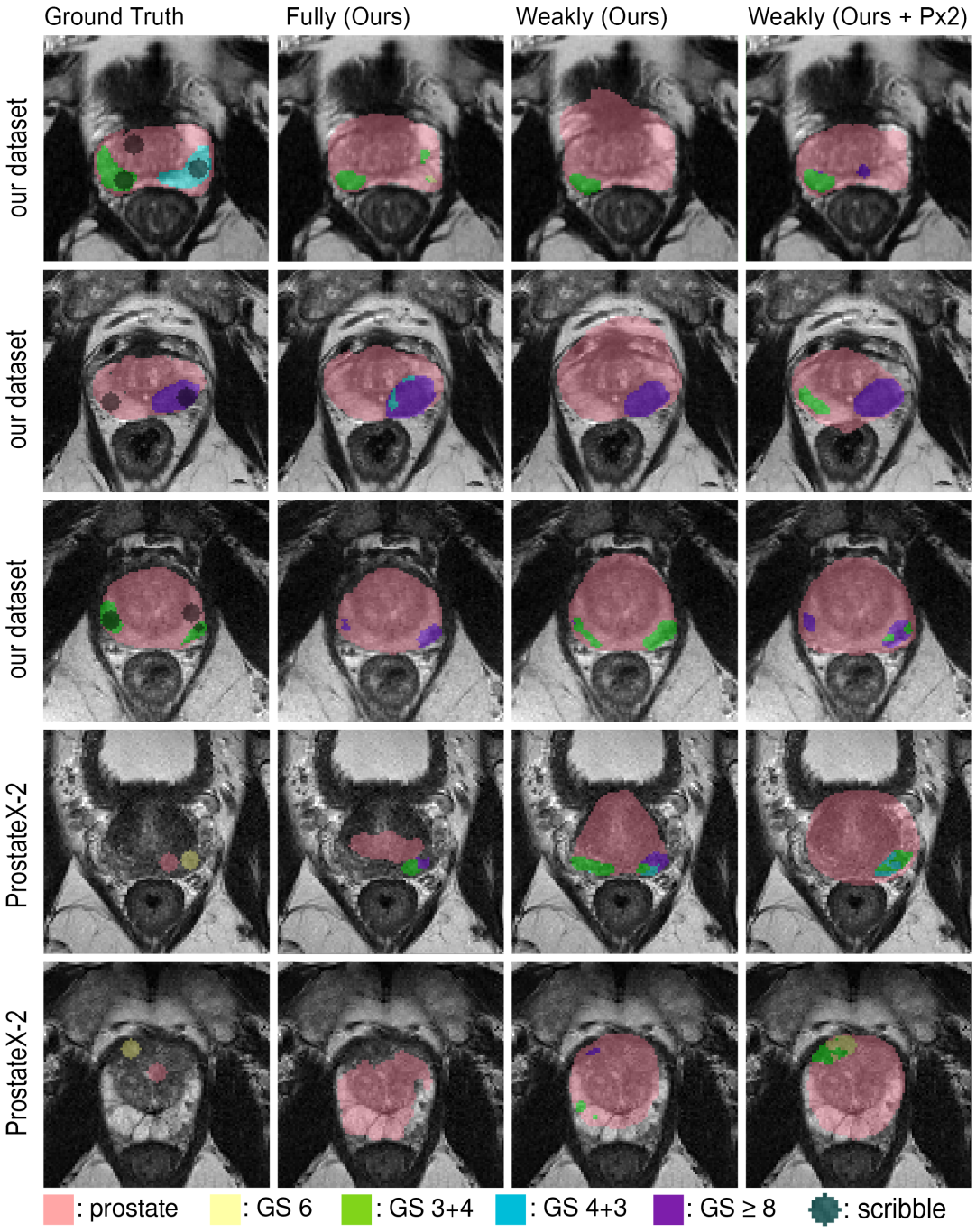


FIGURE 8.10 – Prédiction pour plusieurs images de validation. Les images des trois premières lignes proviennent de CLARA-P, tandis que les images des deux dernières lignes proviennent de ProstateX-2. Deuxième colonne : U-Net entièrement supervisé entraîné avec notre jeu de données. Troisième et quatrième colonnes : U-Net faiblement supervisé entraîné sur les annotations faibles (disques) de CLARA-P seulement et sur les deux jeux de données, respectivement.

Dans le tableau 8.2, nous pouvons voir que les valeurs les plus élevées du coefficient kappa (0.324 ± 0.053) et de la sensibilité à 2 FP (0.649 ± 0.033) sont obtenues avec le modèle de référence totalement supervisé, ce qui est plutôt attendu. Cependant, la performance du modèle U-Net faiblement supervisé, entraîné avec l'EC partielle + *Tags* est proche du modèle entièrement supervisé lorsque l'on considère le coefficient de kappa (0.289 ± 0.072) et la sensibilité à 2 FP (0.587 ± 0.053). Selon le test des rangs signés de Wilcoxon, la différence n'est pas significative avec une p-valeur$= 0.165$ pour le kappa et de 0.409 pour la sensibilité.

Le modèle faiblement supervisé entraîné avec les deux ensembles de données obtient un score de kappa de 0.276 ± 0.037 sur ProstateX-2, ce qui est le score de kappa le plus élevé rapporté dans l'état de l'art sur cette base de challenge pour une tâche de segmentation.

Le modèle faiblement supervisé appris sur ProstateX-2 seulement obtient de très mauvaises performances, comme en témoigne le kappa de -0.002 ± 0.003. Cela peut s'expliquer par la petite taille de l'ensemble de données d'entraînement ProstateX-2 (99 patients), pour lequel seules les coupes présentant au moins une lésion ont été incluses. De plus, les annotations de la prostate saine ont été générées à partir des positions des lésions, ce qui ne permet pas d'inclure des coupes sans lésion, ni sans prostate.

En ce qui concerne les résultats visuels (figure 8.10), nous pouvons voir que même avec des annotations faibles, les cartes de segmentation sont de bonne qualité. La segmentation de la prostate reste moins précise sur CLARA-P avec le modèle faible qu'avec le modèle supervisé, mais la précision de la segmentation s'améliore lorsque les deux jeux de données sont inclus dans l'entraînement. L'ajout de patients issus de la base ProstateX-2 dans l'entraînement améliore d'ailleurs considérablement les performances du modèle sur ce jeu de données. Concernant la généralisation sur le jeu de données ProstateX-2, il est intéressant de noter que le modèle le moins performant est le modèle supervisé. La lésion n'est pas détectée et la prostate est mal segmentée. Il semblerait que le modèle supervisé (entraîné sur les données de CLARA-P uniquement) soit plus enclin au sur-apprentissage et n'apprenne pas la forme de la prostate mais plutôt des caractéristiques concernant chacun des pixels propres à la base CLARA-P, contrairement aux modèles faiblement supervisés.

8.5.3 Discussion

Ces premiers résultats sont prometteurs puisqu'ils montrent des performances qui se rapprochent de la référence totalement supervisée et permettent d'inclure et d'améliorer considérablement les performances sur le jeu de données ProstateX-2.

Toutefois, nous avons généré les disques en ayant connaissance des contours des lésions et fait en sorte qu'ils n'en sortent pas. En pratique, seule une vérité faible sera disponible, il ne sera donc pas possible d'adapter la taille des disques pour ne pas sortir des lésions. Dans les expériences suivantes, nous étudions l'impact d'une taille fixée pour tous les disques ou de disques pouvant sortir de la prostate et des lésions.

8.6 Impact de la taille du disque

8.6.1 Expériences

Dans l'expérience précédente, nous avons considéré comme annotations partielles des disques de rayon 4 pixels, de la même manière que Kervadec et al. [70].

Or, dans les expériences de ce papier, la tâche consiste à segmenter des os ou organes entiers (prostate, ventricule gauche du cœur ou corps vertébral) et non des lésions, de taille inférieure. Dans le cas du CaP, choisir un rayon de 4 pixels peut s'avérer trop grand pour de petites lésions. Dans ces expériences, il s'agit d'évaluer l'impact de la taille du disque sur les performances. Deux scénarios différents sont considérés :

— rayon fixe : le même rayon est utilisé pour toutes les lésions et coupes. Dans ce cas, un disque peut éventuellement sortir de la vérité terrain. En pratique, cela correspond au cas où le radiologue ne pointe que le centre de la lésion, puis le disque est extrapolé avec le rayon choisi.

— rayon variable : comme dans l'expérience précédente, une taille maximale de rayon est établie mais la taille du disque est diminuée si nécessaire de manière à ce qu'il ne sorte pas des contours de la vérité terrain. Cela correspond au cas où le radiologue dessine directement le disque dans la lésion, dont il diminue le rayon si nécessaire.

Quatre rayons différents ont été utilisés pour les deux scénarios, allant de 1 à 4 pixels. Le tableau 8.3 montre le nombre de pixels annotés (ou compris dans le disque) en fonction du rayon considéré. À noter que les mêmes positions de disque (générées de manière aléatoire la première fois) sont utilisées dans toutes les expériences.

rayon	1	2	3	4
nombre de pixels annotés	5	13	29	49

TABLEAU 8.3 – Nombre de pixels annotés en fonction du rayon du disque considéré.

8.6.2 Résultats

rayon max	Kappa	Sensi à 1FP	Sensi à 2FP	Sensi max	Max FP	Dice prostate
1 pixel	0.246 ± 0.058	0.378 ± 0.050	$\mathbf{0.619 \pm 0.041}$	0.819 ± 0.018	$\mathbf{4.05 \pm 0.36}$	0.761 ± 0.006
2 pixels	0.263 ± 0.038	0.352 ± 0.055	0.582 ± 0.056	$\mathbf{0.835 \pm 0.027}$	5.16 ± 0.54	0.759 ± 0.006
3 pixels	$\mathbf{0.314 \pm 0.042}$	0.388 ± 0.081	0.616 ± 0.038	0.816 ± 0.007	4.27 ± 0.63	0.775 ± 0.006
4 pixels	0.289 ± 0.072	$\mathbf{0.416 \pm 0.053}$	0.618 ± 0.044	0.810 ± 0.027	4.51 ± 0.59	$\mathbf{0.800 \pm 0.017}$

TABLEAU 8.4 – Impact de la taille du disque (avec un rayon variable $\leq$ rayon max) sur les performances du modèle faiblement supervisé.

rayon fixe	Kappa	Sensi à 1FP	Sensi à 2FP	Sensi max	Max FP	Dice prostate
1 pixel	0.295 ± 0.050	0.345 ± 0.040	0.578 ± 0.063	0.807 ± 0.030	4.57 ± 0.50	0.819 ± 0.006
2 pixels	0.290 ± 0.034	0.330 ± 0.011	0.566 ± 0.011	$\mathbf{0.836 \pm 0.021}$	4.99 ± 0.40	0.823 ± 0.002
3 pixels	$\mathbf{0.315 \pm 0.015}$	0.384 ± 0.059	0.584 ± 0.050	0.813 ± 0.016	$\mathbf{4.34 \pm 0.25}$	0.823 ± 0.003
4 pixels	0.283 ± 0.044	$\mathbf{0.411 \pm 0.071}$	$\mathbf{0.640 \pm 0.036}$	0.826 ± 0.038	4.76 ± 0.51	$\mathbf{0.825 \pm 0.007}$

TABLEAU 8.5 – Impact de la taille du disque (avec un rayon fixe) sur les performances du modèle faiblement supervisé.

Les résultats sont présentés tableau 8.4 pour un rayon variable. De manière assez étonnante, la meilleure sensibilité à 2FP et le nombre de FP le plus faible sont obtenus avec le plus petit rayon de 1 pixel, correspondant à 5 pixels annotés. Il semblerait que l'apprentissage reste correct avec très peu de données annotées. Toutefois, d'autres

métriques sont plus faibles avec aussi peu de pixels, comme le kappa, la sensibilité à 1 FP ou encore le Dice de la prostate. Les performances générales du modèle restent augmentées lorsque davantage de pixels sont annotés : globalement, les modèles entraînés avec un rayon de 4 pixels montrent les meilleures performances, avec de loin la plus haute sensibilité à 1 FP (0.416 ± 0.053), la meilleure segmentation de la prostate et une sensibilité à 2 FP équivalente au rayon de 1 pixel. Rappelons également que le rayon réel des lésions peut être inférieur au rayon maximal défini, si les lésions sont de petite taille. La faible différence de performances entre les différents cas peut également s'expliquer de cette manière, le rayon maximal défini n'étant pas impactant pour les petites lésions.

Le tableau 8.5 montre les performances pour un rayon fixe, qui peut donc potentiellement sortir de la lésion. En pratique, si les contours des lésions ne sont pas connus, c'est cette approche qui devra être adoptée pour décider du rayon du disque à tracer autour du centre de la lésion. Les résultats sont assez proches du tableau 8.4, avec les meilleures performances observées quand davantage de pixels sont annotés. Cette imprécision, présente dans la vérité terrain, ne semble pas impacter négativement le modèle, le Dice prostate est même augmenté de 5 % par rapport au tableau 8.4 dans tous les cas sauf avec 4 pixels (2.5 % de hausse). On peut expliquer ce faible impact de plusieurs manières :

— les rayons considérés restent faibles et la proportion de pixels mal annotés négligeable ;

— il existe déjà une imprécision dans le tracé de la vérité terrain, qui est potentiellement supérieure à l'imprécision présente ici ;

— les pixels n'appartenant pas aux lésions mais annotés comme tels sont toujours localisés dans la prostate, ce qui peut au contraire aider le réseau à délimiter la prostate.

8.6.3 Discussion

L'utilisation de cette méthode semble faisable en pratique, même sans connaître les contours précis des lésions. Un rayon de taille 4 pixels semble être un choix raisonnable, même si la vérité terrain peut sortir des lésions par moment. Ce rayon s'approche de celui utilisé par Tsehay et al. [125] (approche présentée section 8.2.3), où tous les voxels dans un rayon de 5 mm (soit 5 pixels pour nous) sont considérés comme appartenant à la lésion, hypothèse qui leur permet d'obtenir de bons résultats.

Pour pousser plus loin ce travail, il aurait été intéressant de voir l'impact d'un rayon plus grand, jusqu'à trouver le rayon à partir duquel les performances du modèle régressent.

8.7 Impact de la position du disque

8.7.1 Expériences

Les résultats obtenus dans l'expérience précédente avec l'apprentissage faiblement supervisé sont très satisfaisants, notamment pour la segmentation de la prostate, alors que seulement de petits disques sont annotés. Les disques étant générés de manière aléatoire, les annotations fournies permettent sans doute d'avoir un exemple des différentes parties de la prostate et donc d'apprendre sa représentation. Nous avons voulu vérifier cette hypothèse en générant des annotations présentant

moins de variabilité. De plus, nous avons également étudié l'impact d'annotations imprécises sur l'apprentissage.

Nous avons donc imaginé d'autres méthodes pour générer les annotations faibles :

1. centroïde : le centre de gravité de la lésion ou de la prostate est considéré pour placer le centre du disque. Une limite de cette méthode concerne les lésions non convexes, pour lesquelles le centroïde peut être en dehors de la lésion.
2. plus grand disque : le gribouillis est positionné à l'endroit où le plus grand disque peut rentrer. Une fois la position identifiée, le rayon du disque est réduit jusqu'à être inférieur au rayon maximal fixé.
3. position aléatoire dans l'objet : c'est la méthode utilisée dans les expériences précédentes. Elle consiste à choisir une position aléatoire dans l'objet d'intérêt pour positionner le disque. Si tout le disque ne rentre pas dans l'objet, alors une nouvelle position est choisie. Si après 100 essais aucune position n'a été trouvée pour le disque, son rayon est diminué de 1 pixel et les opérations précédentes réitérées. Cette méthode est appelée *aléatoire valide* et a été utilisée dans les expériences précédentes.
4. position aléatoire flexible : dans ce cas, la position du disque est également aléatoire dans l'objet (lésion ou prostate) mais une imprécision est introduite dans la vérité terrain. Pour introduire cette imprécision, l'objet est dilaté avec un élément structurant carré de taille 5×5 pixels avant d'appliquer la méthode 3 *aléatoire valide*. La position du centre du disque ainsi qu'une partie du disque peuvent donc se retrouver en dehors de l'objet. L'idée est ici de voir l'impact d'une annotation imprécise, étant donné la difficulté de la tâche d'annotation sur les IRM.

8.7.2 Résultats

position	Kappa	Sensi à 1FP	Sensi à 2FP	Sensi max	Max FP	Dice prostate
centroïde	$\mathbf{0.305 \pm 0.052}$	0.351 ± 0.031	0.581 ± 0.049	0.806 ± 0.021	4.93 ± 0.57	0.708 ± 0.005
plus grand disque	0.280 ± 0.011	0.417 ± 0.024	0.640 ± 0.018	$\mathbf{0.828 \pm 0.015}$	5.14 ± 0.38	0.749 ± 0.012
aléatoire flexible	0.274 ± 0.043	$\mathbf{0.445 \pm 0.070}$	$\mathbf{0.657 \pm 0.051}$	0.811 ± 0.010	$\mathbf{4.21 \pm 0.85}$	0.796 ± 0.004
aléatoire valide	0.289 ± 0.072	0.416 ± 0.053	0.618 ± 0.044	0.810 ± 0.027	4.51 ± 0.59	$\mathbf{0.800 \pm 0.017}$

TABLEAU 8.6 – Impact de la position du disque sur les performances du modèle faiblement supervisé.

Le tableau 8.6 montre les résultats pour les différentes positions testées. À noter que *aléatoire valide* correspond au type d'annotation utilisé dans la section 8.5 et pour l'expérience avec rayon variable de la section 8.6 (les résultats pour cette méthode sont les mêmes que dans le tableau 8.4 avec un rayon de 4 pixels). Les résultats les moins bons sont observés avec la méthode *centroïde* (hormis pour le kappa). Comme illustré figure 8.11, cette méthode présente déjà des inconvénients pour les formes non convexes, pour lesquelles le centre de gravité n'est pas forcément inclus dans l'objet. De plus, le réseau n'est pas entraîné à reconnaître les bords puisque seul le centroïde est annoté. Cela se reflète particulièrement dans la métrique Dice pour la prostate, qui est très faible avec cette méthode. Les prostates segmentées sont ainsi toujours bien plus petites que la référence (voir figure 8.12). Au contraire, en utilisant une méthode aléatoire, les disques sont placés uniformément dans toutes les régions de la prostate ; cette hétérogénéité permet au modèle d'apprendre plus précisément

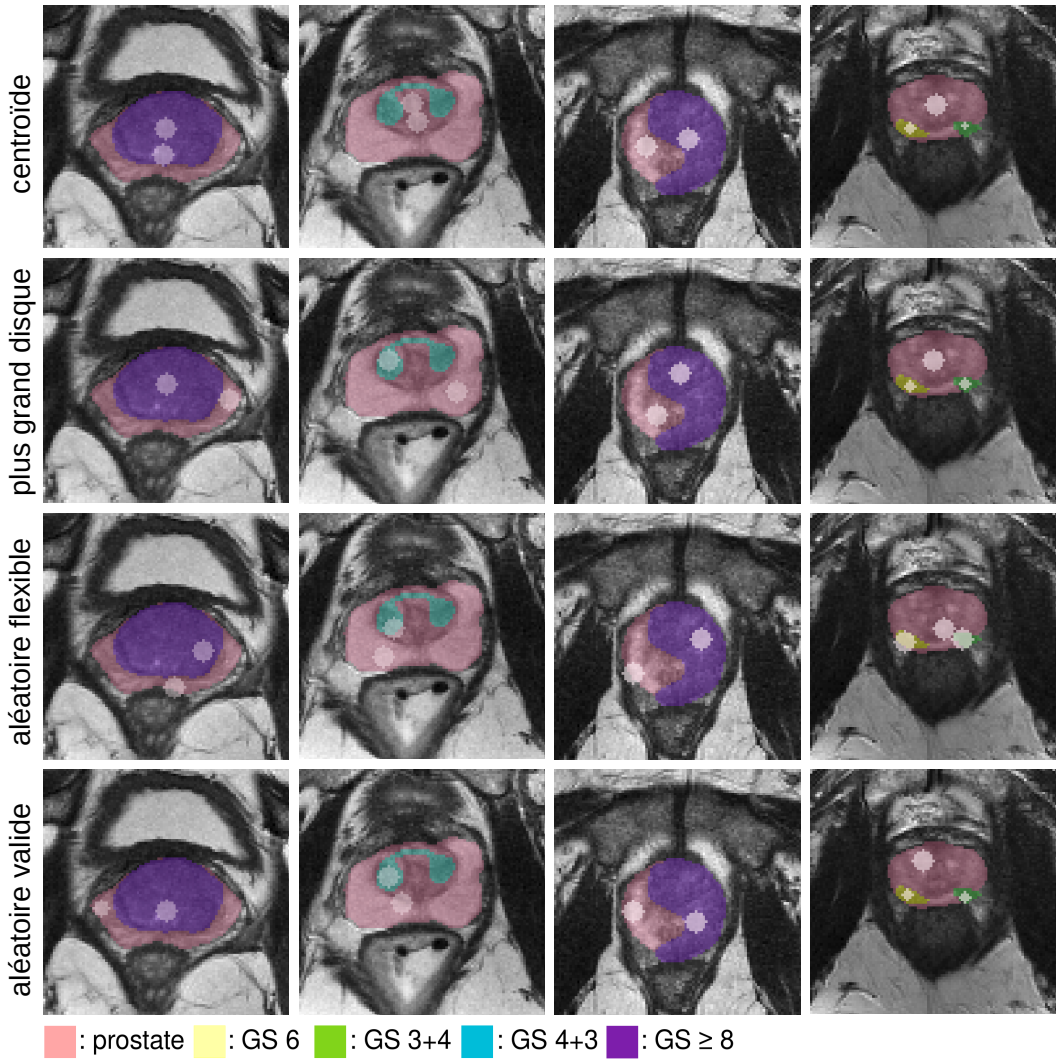


FIGURE 8.11 – Annotations faibles (disques) générées selon les différentes méthodes testées. Le rayon maximal utilisé ici est de 4 pixels. Les deux premières colonnes montrent des exemples où la méthode du centroïde échoue : dans le premier cas, le disque correspondant à la prostate se trouve dans la lésion et inversement pour le second cas.

ce qu'est la prostate et de produire de meilleures cartes de segmentation. La méthode du *plus grand disque* se rapproche de celle du *centroïde*, avec un Dice prostate bien plus faible que les deux autres méthodes et une variabilité dans les disques de la prostate bien plus faible. Nous pouvons observer que la méthode *aléatoire flexible* présente de meilleures performances que la méthode *aléatoire valide* (excepté pour le kappa et le Dice prostate), bien que très proches. Cela confirme que la présence de quelques voxels mal annotés n'altère pas l'apprentissage.

La figure 8.12 montre des prédictions visuelles pour chacune des méthodes, avec un exemple pour chaque scanner. On observe effectivement des prostates sous-estimées avec les deux premières méthodes et surestimées avec les deux méthodes aléatoires. Cette figure permet également de rendre compte du nombre croissant de lésions prédites pour les modèles apprenant avec des annotations qui sont moins représentatives de l'hétérogénéité présente dans les données. Cela est particulièrement visible sur le dernier exemple (scanner Philips), où une grande partie de la prostate est prédite comme touchée par le CaP.

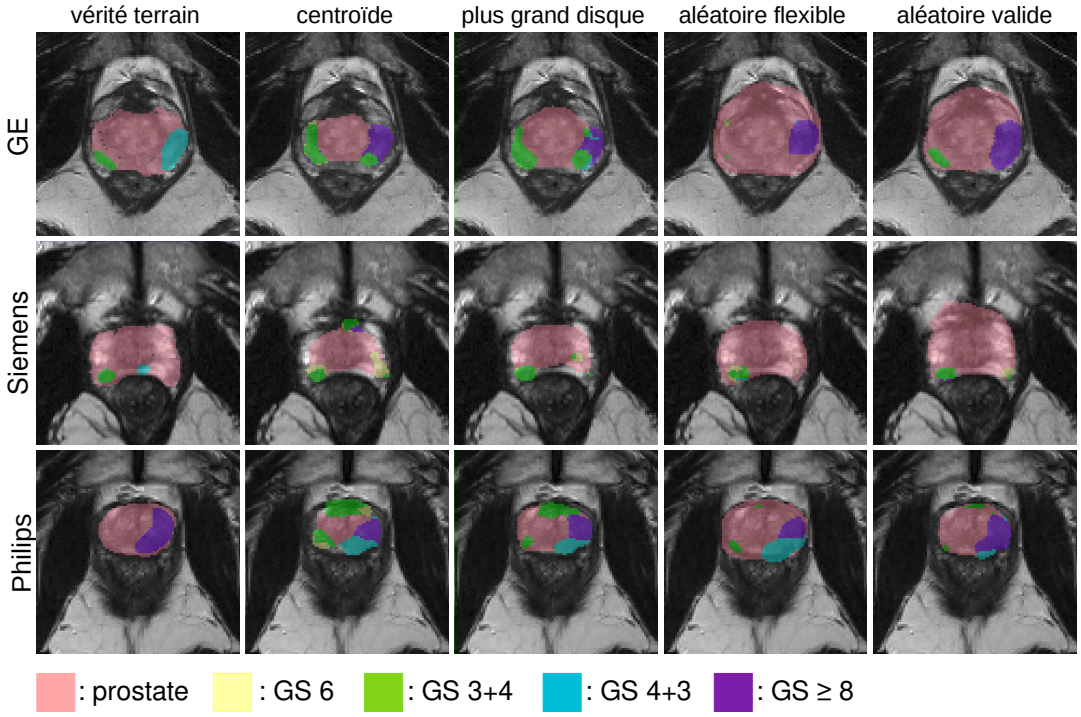


FIGURE 8.12 – Exemple de prédictions selon les différentes manières de générer les annotations faibles.

8.7.3 Discussion

Ces expériences nous ont permis de valider que les bons résultats observés étaient en partie dus à la variabilité présente dans la génération des annotations faibles. Des disques localisés toujours au même endroit (centroïde notamment) ne permettent pas au modèle d'apprendre les délimitations de la prostate. Ces résultats mettent également en avant la problématique des métriques, qui ne traduisent pas toutes les informations (comme la taille des lésions prédites par exemple). Une inspection visuelle des résultats reste indispensable à toute conclusion.

8.8 Ajout de contraintes plus fines sur la taille des prédictions

8.8.1 Expériences

Les expériences précédentes utilisent une contrainte large sur la taille des prédictions (voir équation 8.3), où la taille de la prédiction doit être ≥ 1 si un objet de la classe en question est présent dans l'image et nulle dans le cas contraire (d'où le nom de *Tags*). Cette contrainte pourrait être plus fine, comme présenté dans le papier de Kervadec et al. [70], avec des contraintes appelées *Common Bounds* (noté CB) communes à tous les patients. Nous avons étudié l'impact de contraintes plus précises qui seraient issues de la distribution observée des tailles de lésion et prostate dans le jeu de données CLARA-P. La contrainte étant appliquée sur des coupes 2D, nous avons étudié les distributions en 2D pour les 219 patients inclus dans l'étude. Les statistiques principales sont présentées tableau 8.7.

Comme observé dans ce tableau, la taille des lésions est corrélée à l'agressivité du cancer. Nous avons donc considéré des bornes distinctes pour chacune des classes

	prostate	**GS 3+3**	**GS 3+4**	**GS 4+3**	**GS $\geq$ 8**
taille min	17	6	5	8	9
taille max	3475	487	610	817	1059
taille moyenne	996	83	108	142	220

TABLEAU 8.7 – Statistiques concernant la distribution des tailles de la prostate et des lésions en 2D pour les 219 patients de la base CLARA-P inclus dans l'étude.

de lésions, définies tableau 8.8. Le choix n'est pas évident compte tenu de la variabilité des tailles de lésion au sein même d'une classe. Nous avons opté pour une borne inférieure a commune à toutes les lésions et égale à la plus petite taille de lésion observée, soit 5 pixels. C'est donc surtout la borne supérieure b qui sera impactante lors de l'apprentissage. Nous avons choisi des bornes b supérieures à la taille moyenne, mais inférieures à la taille maximale observée, excluant ainsi quelques *outliers*. Il en est de même pour la prostate, où les plus grosses prostates ne sont pas comprises dans l'intervalle fixé. Une borne supérieure b trop grande serait sans effet, nous menant à choisir malgré tout une contrainte qui n'est pas respectée pour certaines prostates.

	prostate	**GS 3+3**	**GS 3+4**	**GS 4+3**	**GS $\geq$ 8**
a (borne inférieure)	100	5	5	5	5
b (borne supérieure)	2500	300	350	450	800

TABLEAU 8.8 – Choix des bornes a et b définies équation 8.3 pour chacune des classes.

Les résultats sont évalués comme précédemment : la prostate et les lésions sont annotés par des disques de rayon 4 pixels, la méthode *aléatoire valide* est utilisée pour générer les annotations faibles et 4 réplicats de validation croisée à 5 plis sont réalisés. La contrainte CB est ajoutée à l'entropie croisée partielle ainsi qu'à la contrainte des *Tags*, aboutissant à la fonction de coût suivante :

$$\mathcal{H}(S) + \lambda \mathcal{C}_{Tags}(V_S) + \lambda \mathcal{C}_{CB}(V_S) \tag{8.5}$$

avec λ toujours fixé à 10^{-5} pour chacune des contraintes.

8.8.2 Résultats

Méthode	Kappa	Sensi à 1FP	Sensi à 2 FP	Sensi max	Max FP	Dice prostate
Totalement supervisé	$\mathbf{0.324 \pm 0.053}$	$\mathbf{0.541 \pm 0.062}$	0.649 ± 0.033	0.672 ± 0.014	$\mathbf{2.10 \pm 0.66}$	0.799 ± 0.004
EC part. + Tags	0.289 ± 0.072	0.416 ± 0.053	0.618 ± 0.044	0.810 ± 0.027	4.51 ± 0.59	$\mathbf{0.800 \pm 0.017}$
EC part. + Tags + CB	0.305 ± 0.034	0.459 ± 0.045	$\mathbf{0.699 \pm 0.028}$	$\mathbf{0.828 \pm 0.020}$	3.62 ± 0.27	0.777 ± 0.01

TABLEAU 8.9 – Apport de la contrainte *Common Bounds* (CB) dans l'apprentissage faiblement supervisé évalué sur 4 réplicats de validation croisée à 5 plis. EC part. : entropie croisée partielle.

Les résultats concernant l'ajout d'une contrainte plus précise et dépendant de la classe sont présentés tableau 8.9. Ils sont comparés à la référence totalement supervisée et à la première approche faiblement supervisée avec une contrainte *Tags* déjà présentées tableau 8.2. L'ajout de cette contrainte plus forte est bénéfique pour toutes les métriques rapportées dans le tableau, sauf le Dice prostate qui est étonnamment un peu plus faible. Cela peut s'expliquer par la pénalité pouvant être très

importante si une prostate trop grande est prédite, puisqu'elle comprend bien plus de pixels qu'une lésion et peut donc mener à des quantités $(V_S - b)^2$ très grandes et pénalisantes.

Certaines métriques sont même meilleures que la référence totalement supervisée comme la sensibilité à 2 FP (0.699 ± 0.028 vs. 0.649 ± 0.033 pour la référence) et la sensibilité maximale. Toutefois, le nombre de FP maximal moyen est bien plus élevé que la référence totalement supervisée ; la meilleure sensibilité finale se fait au prix de davantage de lésions prédites, dont certaines peuvent intersecter la vérité terrain peut-être par hasard.

8.8.3 Discussion

L'ajout de cette contrainte sur la taille des lésions semble bénéfique au modèle, mais ces résultats restent préliminaires. Il serait intéressant d'étudier plus finement l'influence de chaque contrainte dans l'apprentissage et d'optimiser davantage le choix des bornes pour chacune des lésions. La borne supérieure de la classe prostate pourrait également être augmentée à la plus grande taille de prostate observée, pour éviter que le modèle ne prenne le risque de prédire de trop grandes prostates au prix d'un Dice final inférieur à la référence sans contrainte CB.

Par ailleurs, cette expérience permet également de mettre en lumière la problématique liées aux métriques : des valeurs de sensibilité plus élevées à un certain taux de FP ne traduisent pas forcément un meilleur modèle mais un modèle qui prédit davantage de lésions. Reste à savoir s'il est préférable pour un radiologue d'avoir un modèle avec une meilleure sensibilité finale, mais également avec davantage de FP, ou des prédictions peut-être plus fiables mais moyennant une sensibilité finale plus faible...

8.9 Conclusion

Les résultats montrent des performances équivalentes à celles obtenues avec une approche totalement supervisée en entraînant un modèle U-Net avec seulement 6.35 % des pixels annotés grâce à une fonction de coût combinant l'entropie croisée partielle et une contrainte de taille dérivée de celle de Kervadec et al. [70]. Ces résultats sont particulièrement intéressants si l'on considère le gain de temps que représente l'utilisation de gribouillis au lieu d'une annotation au niveau du pixel. En outre, une annotation basée sur un point est très pertinente pour la caractérisation du CaP, la plupart des bases de données reposant sur des résultats de biopsie. Ces modèles de segmentation basés sur des annotations faibles sont susceptibles de faciliter l'inclusion de données provenant de différentes sources (par exemple, différents centres ou scanners) et aider à résoudre le problème récurrent du changement de domaine dans la segmentation d'images médicales. Les perspectives seraient d'étudier un entraînement hybride combinant des données entièrement et faiblement annotées. Il pourrait également être intéressant d'étudier l'utilisation de contraintes de forme sur les lésions et la prostate, adaptées à notre application. Enfin, d'autres méthodes basées sur des annotations faibles telles que décrites section 8.2 pourraient être comparées à cette première approche faiblement supervisée.

Chapitre 9

Apport de l'imagerie dynamique pour la segmentation du cancer de la prostate par agressivité

Ce chapitre est adapté de Audrey Duran, Gaspard Dussert et Carole Lartizien. « Perfusion Imaging in Deep Prostate Cancer Detection from MP-MRI : Can We Take Advantage of it ? » In : 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). *2022, p. 1-5.* DOI *: 10. 1109/ ISBI52829. 2022. 9761616.*

Sommaire

9.1 Introduction

Comme vu dans le chapitre 4, de nombreux systèmes CAD pour la détection et la segmentation du cancer de la prostate (CaP) à partir d'IRM sont basés sur des réseaux de neurones convolutifs (CNN). Ces architectures rendent difficile l'inclusion de données en 4 dimensions (4D) telles que la séquence dite dynamique, ou de perfusion (décrite section 2.7.4). L'impact de cette séquence d'imagerie basée sur l'injection d'un agent de contraste au Gadolinium (DCE) dans la détection du CaP s'avère controversé [128, 28, 134, 9]. Par conséquent, à notre connaissance, la totalité des modèles de segmentation profonds pour le CaP utilisant les images IRM en entrée (excluant ainsi les approches radiomiques) n'incluent que les séquences T2-w et DWI (dont les cartes ADC), conduisant à des modèles biparamétriques (bp). Quelques modèles de classification du CaP incluent la séquence DCE avec les cartes paramétriques K^{trans} issues de la modélisation pharmacocinétique bicompartimentale de Tofts [4, 82]. Dans cet article, nous nous interrogeons sur la contribution de la séquence dynamique dans le contexte de la segmentation et de la caractérisation du CaP à partir d'IRM-mp.

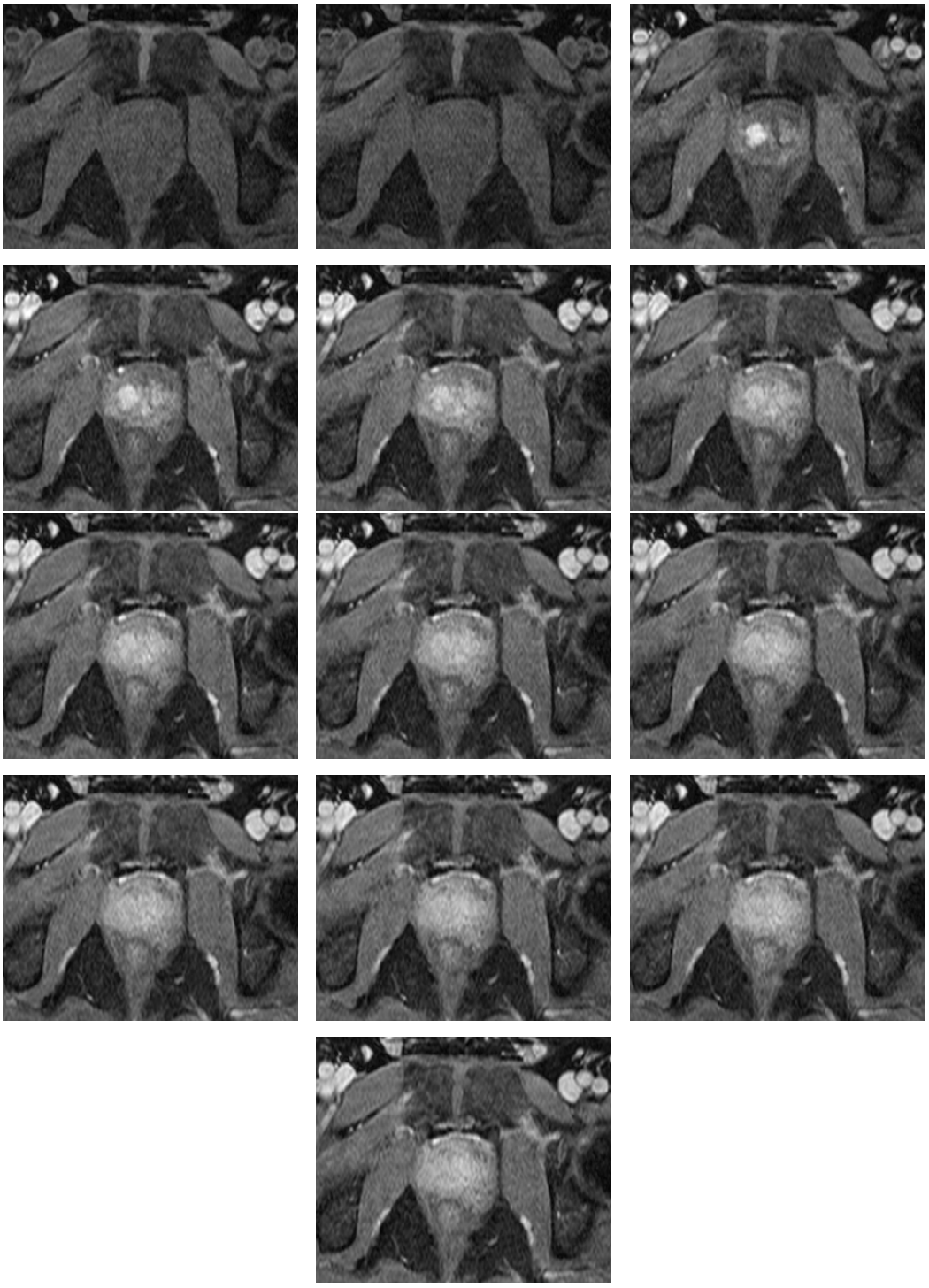


FIGURE 9.1 – Images correspondant aux 13 temps d'une séquence IRM de perfusion chez un patient Siemens. Une lésion est présente et visible en hypersignal sur la 3ème image (temps=3). C'est ce même patient qui est utilisé pour la figure 9.2 et la figure 9.3.

9.2 Matériel et méthodes

9.2.1 Cartes issues de la perfusion

Plusieurs cartes de perfusion (illustrées par exemple sur la figure 9.3) ont été dérivées des séries d'images temporelles (3D + temps) acquises pour chaque patient après l'injection en bolus d'un agent de contraste au gadolinium (voir détails d'acquisition section 5.2 et un exemple 2D + temps figure 9.1). Ces cartes consistent soit en des volumes IRM 3D extraits à des points spécifiques de la série temporelle, soit en des cartes paramétriques semi-quantitatives extraites du traitement des courbes cinétiques au niveau du voxel. Ces paramètres semi-quantitatifs ont par ailleurs été décrits section 2.7.4 et illustrés pour une région d'intérêt suspecte figure 2.12.

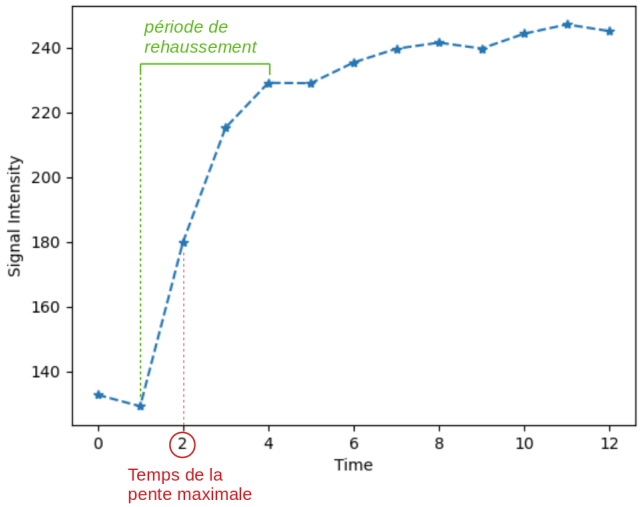
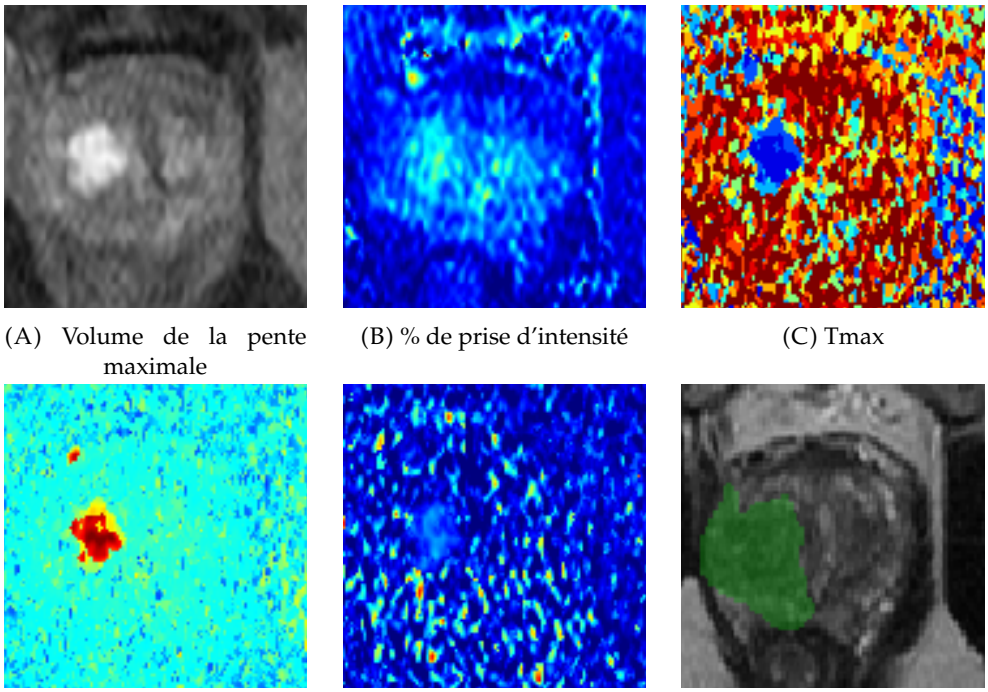


FIGURE 9.2 – Variation de l'intensité moyenne du signal dans le champ de vue en fonction du temps. Cette courbe permet d'extraire le volume 3D correspondant au temps de pente maximale (figure 9.3A) et de définir la période de rehaussement utilisée dans les cartes du % de prise d'intensité (figure 9.3B). Le patient Siemens dont la courbe est issue est le même que celui présenté figure 9.3.

Tmax La carte des *Tmax* est une carte paramétrique, où la valeur de chacun des voxels correspond au temps où l'intensité maximale a été observée pour ce voxel au cours de l'acquisition temporelle. Les zones suspectes correspondent à un temps court, puisqu'elles sont pour la plupart hypervascularisées. L'unité de temps est variable comme la séquence DCE présente une résolution temporelle variable selon les scanners et patients.

Vitesse de rehaussement (ou *wash-in*) La période de rehaussement correspond à la période allant de l'arrivée du gadolinium jusqu'au pic d'intensité, le temps d'arrivée étant défini ici comme le temps correspondant à l'accélération (dérivée seconde) maximale sur la courbe d'intensité temporelle du voxel (voir exemple à l'échelle de l'image figure 9.2) et le pic d'intensité le temps où l'intensité maximale est observée. Les cartes de vitesse de rehaussement sont obtenues en calculant la pente observée dans la courbe temps-intensité pendant la période de rehaussement au niveau du voxel. Plus la pente de rehaussement est élevée, plus la vitesse de rehaussement est rapide, ce qui est un signe suspect de tissu cancéreux.



(A) Volume de la pente maximale (B) % de prise d'intensité (C) Tmax

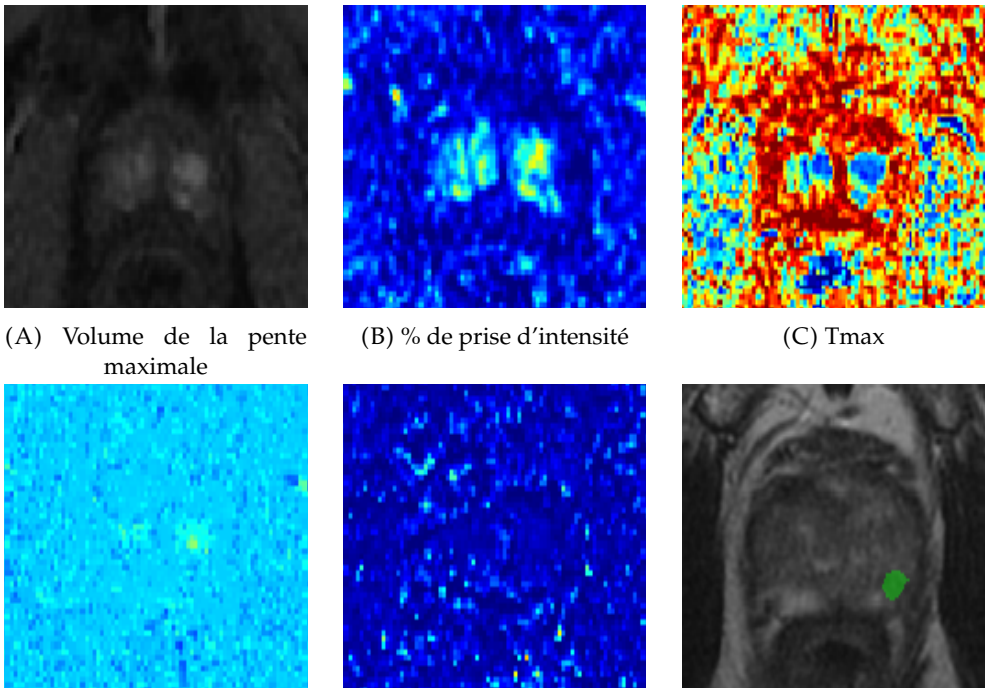
(D) pente du wash-in (E) pente du wash-out (F) T2-w et vérité terrain

FIGURE 9.3 – Exemples des cartes de perfusion considérées dans cette étude pour un patient Siemens Symphony. Les valeurs les plus élevées des cartes paramétriques ((B) à (E)) sont représentées en rouge tandis que les valeurs les plus faibles sont en bleu. Cet exemple montre une lésion GS 3+4 dessinée sur la séquence T2-w.

La vitesse de lavage (ou *wash-out*) Tout comme la carte de rehaussement, la carte des vitesses de lavage contient les pentes observées dans la courbe temps-intensité de chaque voxel pendant la période de lavage allant de Tmax (décrit ci-dessus, variable selon le voxel considéré) à la fin de l'acquisition. Des zones suspectes montrent une vitesse de lavage rapide après le pic maximal de contraste.

Volume du temps de pente maximale Cette carte correspond au volume 3D de la série temporelle DCE où la pente (ou prise d'intensité) maximale a été observée dans la courbe temps-intensité moyenne. Cette courbe est obtenue en calculant la moyenne des intensités dans chaque volume 3D acquis à chacun des temps. Sur l'exemple de courbe temps-intensité figure 9.2, le volume extrait est celui acquis au temps $t = 2$, la pente maximale observée étant entre les temps $t = 1$ et $t = 2$.

Pourcentage maximal des cartes de rehaussement Cette carte 3D reflète le pourcentage maximal de rehaussement entre la première image de l'acquisition et toutes celles obtenues pendant la période de lavage, quel que soit le moment où le rehaussement maximal apparaît. La période de lavage est définie à partir de la courbe temps-intensité. Sur l'exemple figure 9.2, la fin de la période de rehaussement correspond au temps $t = 4$, temps précédent la plus grande décélération observée. Ces cartes ont été décrites dans [138], où la fin de la période de rehaussement était définie à $t = 1$ minute. Les zones suspectes sont en hyper-signal, ou autrement dit présentent de fortes intensités.



(A) Volume de la pente maximale (B) % de prise d'intensité (C) Tmax

(D) pente du wash-in (E) pente du wash-out (F) T2-w et vérité terrain

FIGURE 9.4 – Exemples des cartes de perfusion considérées dans cette étude pour un patient GE Discovery. Les valeurs les plus élevées des cartes paramétriques ((b) à (e)) sont représentées en rouge tandis que les valeurs les plus faibles sont en bleu. Cet exemple montre une lésion GS 3+4 dessinée sur la séquence T2-w.

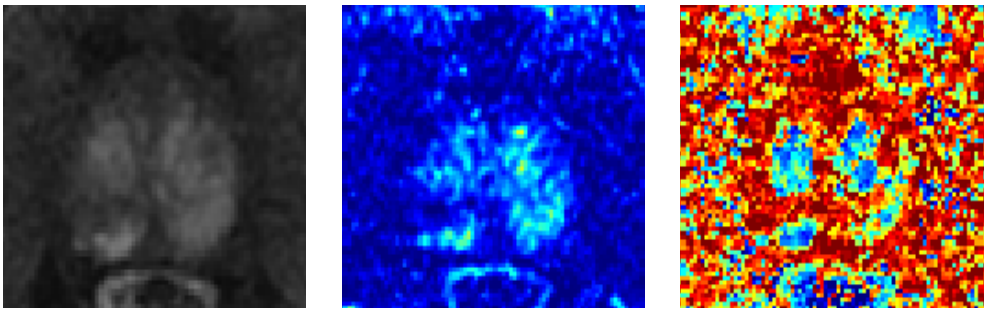
Les cartes pharmacocinétiques quantitatives issues de la modélisation compartimentale telles que K^{trans}, K_{ep}, V_e et V_p n'ont pas été prises en compte en raison de leur forte dépendance à la fonction d'entrée artérielle et de leur variabilité en fonction des paramètres d'acquisition du scanner [11].

9.2.2 Modèle de segmentation et caractérisation du CaP par GS

Le modèle utilisé dans ce travail est basé sur un U-Net standard à quatre blocs, avec des couches de normalisation par lots pour réduire le sur-apprentissage et des activations *Leaky ReLU*. Il produit des cartes de segmentation à 6 canaux, correspondant à 6 étiquettes de classe : pour le fond, la zone globale de la prostate, les lésions GS 6, GS 3+4, GS 4+3 et GS $\geq$ 8. Cette architecture U-Net standard s'est avérée efficace pour la détection et la caractérisation du CaP à partir d'IRM-bp (voir section 7.4.2). Deux stratégies de fusion des différentes modalités ont été envisagées dans ce travail :

— **fusion précoce**, où toutes les modalités considérées sont concaténées en entrée du réseau dans différents canaux ;

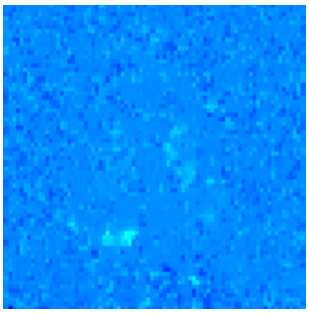
— **fusion intermédiaire**, où chaque modalité est encodée indépendamment dans des branches distinctes du U-Net, jusqu'à une fusion des cartes de caractéristiques dans l'espace latent. Une branche de décodage commune est donc partagée par les modalités.



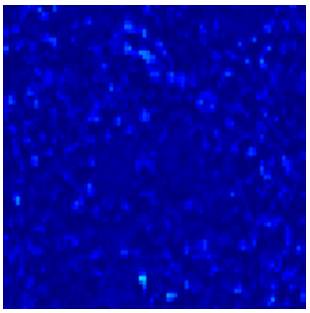
(A) Volume de la pente maximale

(B) % de prise d'intensité

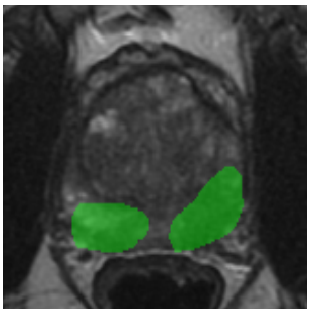
(C) Tmax



(D) pente du wash-in



(E) pente du wash-out



(F) T2-w et vérité terrain

FIGURE 9.5 – Exemples des cartes de perfusion considérées dans cette étude pour un patient GE Discovery. Les valeurs les plus élevées des cartes paramétriques ((b) à (e)) sont représentées en rouge tandis que les valeurs les plus faibles sont en bleu. Cet exemple montre deux lésions GS 3+4 dessinées sur la séquence T2-w.

Pour chaque stratégie de fusion, nous avons considéré un modèle biparamétrique incluant les cartes T2-w et ADC ainsi que différents modèles multiparamétriques prenant en compte l'une des cartes de perfusion listées ci-dessus en plus des images T2-w et ADC.

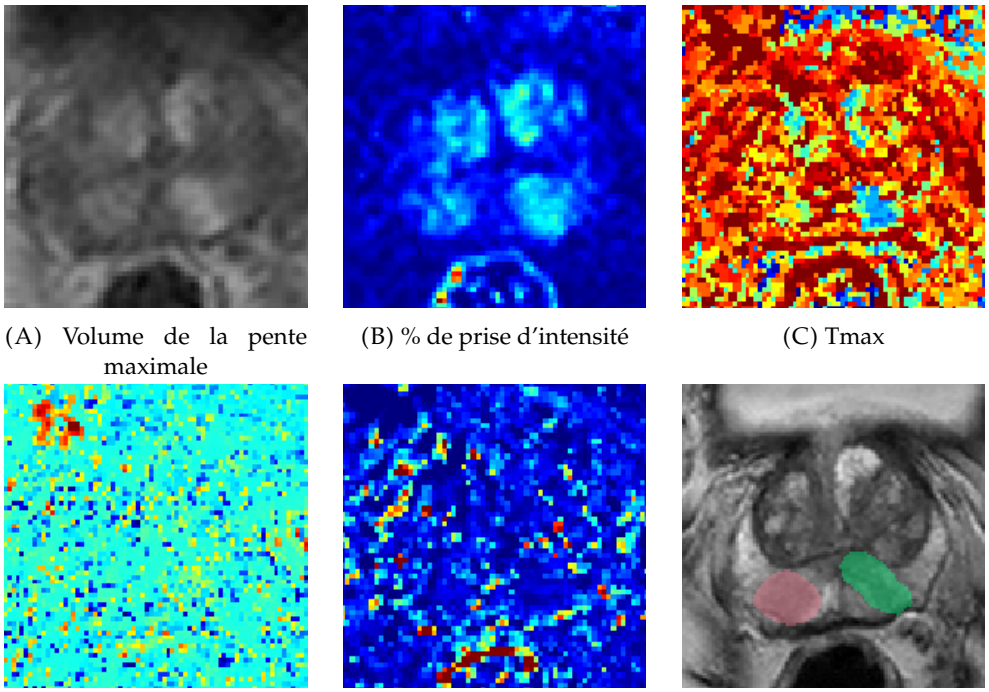
Détails expérimentaux

Les images T2-w et ADC ont été prétraitées comme présenté section 6.3.1.4 : l'intensité a été normalisée dans l'intervalle [0, 1], les volumes sont ré-échantillonnés à une résolution de $1 \times 1 \times 3$ mm^3 et chaque coupe est automatiquement recadrée à une taille de 96×96 pixels au centre de l'image.

Chaque configuration a été évaluée en validation croisée à 5 plis, en faisant la moyenne de 4 réplicats de validation croisée pour chaque expérience compte tenu de la variabilité observée entre plusieurs expériences. Les patients ont été répartis dans les 5 plis de manière à équilibrer autant que possible le nombre de lésions par classe et le nombre de patients de chaque scanner. La répartition des patients CLARA-P est la même que dans le chapitre 7 et le 8.

Pendant la phase d'apprentissage, des techniques d'augmentation de données classiques (rotation, zoom, décalage, retournement ou *flip* horizontal) ont été appliquées afin de réduire le risque de sur-apprentissage.

Tous les réseaux ont été optimisés en utilisant Adam et une régularisation L2 avec $\gamma = 10^{-5}$. Le taux d'apprentissage initial a été fixé à 10^{-3} avec une décroissance de 0.5 après 25 époques sans amélioration de la fonction de coût en validation. Après



(A) Volume de la pente maximale (B) % de prise d'intensité (C) Tmax

(D) pente du wash-in (E) pente du wash-out (F) T2-w et vérité terrain

FIGURE 9.6 – Exemples des cartes de perfusion considérées dans cette étude pour un patient Philips Ingenia. Les valeurs les plus élevées des cartes paramétriques ((b) à (e)) sont représentées en rouge tandis que les valeurs les plus faibles sont en bleu. Cet exemple montre une lésion GS 3+4 dessinée en vert sur la séquence T2-w et une zone suspecte mais saine en rose.

50 époques sans amélioration, l'entraînement est arrêté (*early stopping*). Les lésions de taille inférieure à 15 voxels (c'est-à-dire 45 mm^3) ont été supprimées.

Dans ce chapitre, la librairie Keras-TensorFlow 2.4 est utilisée. Les métriques de performances sont les mêmes que dans les deux chapitres précédents, décrites section 6.3.4.

9.3 Résultats

9.3.1 Performance de détection

Performance de détection binaire Le tableau 9.1 montre les performances de détection et de segmentation des lésions CS (GS $>$ 6) pour chacune des cartes de perfusion en fusion précoce ou intermédiaire par rapport au modèle de base IRM-bp. En ce qui concerne la stratégie de fusion précoce, l'ajout du volume *pente maximale*, de la carte *Tmax* ou *wash-out* aux images T2-w et ADC permet d'améliorer les performances du modèle par rapport au modèle de base. Pour les cartes *pente maximale* et *Tmax*, nous observons un gain de sensibilité de 5.4 % et 4.1 % respectivement à 1 FP, 2.6 % et 3.3 % à 2 FP et la sensibilité maximale augmente de 5.8 % et 5.9 %. Pour ces 2 modèles, le nombre maximal de FP est également réduit de 0.152 pour la *pente maximale* et de 0.172 pour le *Tmax*. Les cartes de *% de rehaussement* et de *wash-in* n'apportent pas d'informations discriminantes pour cette tâche. En considérant le coefficient kappa de Cohen, la meilleure valeur (0.343 ± 0.050) est obtenue avec le modèle *wash-out*, qui surpasse la référence (0.318 ± 0.019).

Modèle	Kappa	Sensi à 1FP	Sensi à 2FP	Sensi max	Max FP	Dice prostate
Fusion précoce						
référence : IRM-bp	0.318 ± 0.019	0.544 ± 0.029	0.660 ± 0.030	0.674 ± 0.031	2.134 ± 0.139	$\mathbf{0.789 \pm 0.002}$
IRM-bp + pente max.	0.303 ± 0.054	$\mathbf{0.598 \pm 0.026}$	0.686 ± 0.036	0.705 ± 0.039	1.982 ± 0.304	0.769 ± 0.007
IRM-bp + % rehau.	0.312 ± 0.035	0.525 ± 0.019	0.649 ± 0.025	0.693 ± 0.030	2.525 ± 0.185	0.771 ± 0.002
IRM-bp + Tmax	0.328 ± 0.026	0.585 ± 0.020	$\mathbf{0.693 \pm 0.024}$	$\mathbf{0.706 \pm 0.019}$	$\mathbf{1.962 \pm 0.165}$	0.770 ± 0.008
IRM-bp + wash-in	0.281 ± 0.020	0.541 ± 0.041	0.659 ± 0.030	0.687 ± 0.024	2.308 ± 0.388	0.784 ± 0.013
IRM-bp + wash-out	$\mathbf{0.343 \pm 0.050}$	0.553 ± 0.021	0.665 ± 0.015	0.680 ± 0.015	2.302 ± 0.104	0.778 ± 0.003
IRM-bp + Tmax + pente max.	0.295 ± 0.023	0.569 ± 0.040	0.664 ± 0.021	0.697 ± 0.026	2.066 ± 0.453	0.759 ± 0.013
Fusion intermédiaire						
référence : IRM-bp	0.333 ± 0.060	0.529 ± 0.017	0.656 ± 0.010	0.687 ± 0.010	2.440 ± 0.341	0.792 ± 0.004
IRM-bp + pente max.	$\mathbf{0.378 \pm 0.033}$	0.569 ± 0.029	0.693 ± 0.022	0.708 ± 0.017	$\mathbf{1.929 \pm 0.111}$	$\mathbf{0.798 \pm 0.004}$
IRM-bp + Tmax	0.315 ± 0.064	$\mathbf{0.582 \pm 0.041}$	$\mathbf{0.712 \pm 0.009}$	$\mathbf{0.732 \pm 0.007}$	2.331 ± 0.399	0.778 ± 0.008

TABLEAU 9.1 – Performance de détection et segmentation des lésions. Les résultats correspondent à la moyenne des métriques obtenues sur 4 réplicats de la validation croisée à 5 plis. Les meilleurs résultats pour chaque métrique et stratégie de fusion sont en gras.

Les performances obtenues avec les deux cartes paramétriques les plus performantes (*Tmax* et volume de la *pente maximale*) du scénario de fusion précoce ont été comparées à celles obtenues avec les mêmes cartes de perfusion dans la stratégie de fusion intermédiaire. Cette approche ne montre pas un impact clair : à 1FP, les sensibilités de ces trois modèles (deuxième partie du tableau 9.1) sont plus faibles qu'avec la stratégie de fusion précoce, mais à 2 FP, les sensibilités de *pente maximale* et *Tmax* sont plus élevées qu'avec la stratégie de fusion précoce.

Performance de détection multiclasse Le tableau 9.2 et le tableau 9.3 permettent une analyse plus fine de la sensibilité pour chaque groupe de GS (à 1 et 1.5 FP respectivement). Ils reflètent la capacité du modèle à la fois à localiser les lésions et à leur attribuer le bon grade.

Modèle	GS ≥ 8	GS 4+3	GS 3+4	GS 3+3
Fusion précoce				
référence : IRM-bp	0.56 ± 0.05	0.41 ± 0.06	$\mathbf{0.37 \pm 0.01}$	0.12 ± 0.04
IRM-bp + pente max.	$\mathbf{0.62 \pm 0.02}$	0.45 ± 0.04	0.37 ± 0.02	$\mathbf{0.15 \pm 0.03}$
IRM-bp + % rehau.	0.61 ± 0.02	0.40 ± 0.06	0.36 ± 0.04	0.06 ± 0.03
IRM-bp + Tmax	0.61 ± 0.03	0.46 ± 0.01	0.36 ± 0.01	0.08 ± 0.03
IRM-bp + wash-in.	0.52 ± 0.05	0.45 ± 0.02	0.37 ± 0.05	0.09 ± 0.02
IRM-bp + wash-out	0.59 ± 0.05	0.45 ± 0.03	0.37 ± 0.03	0.14 ± 0.02
IRM-bp + Tmax + pente max.	0.58 ± 0.04	$\mathbf{0.49 \pm 0.04}$	0.35 ± 0.05	0.12 ± 0.03
Fusion intermédiaire				
référence : IRM-bp	0.55 ± 0.03	0.44 ± 0.09	0.38 ± 0.04	$\mathbf{0.17 \pm 0.02}$
IRM-bp + pente max.	$\mathbf{0.56 \pm 0.03}$	0.37 ± 0.03	0.39 ± 0.02	0.16 ± 0.04
IRM-bp + Tmax	0.54 ± 0.03	$\mathbf{0.49 \pm 0.04}$	$\mathbf{0.43 \pm 0.06}$	0.10 ± 0.02

TABLEAU 9.2 – Sensibilité de détection moyenne pour chaque groupe de GS à 1 FP. Les résultats correspondent à la moyenne des métriques obtenues sur 4 réplicats de la validation croisée à 5 plis. Les meilleurs résultats pour chaque métrique et stratégie de fusion sont en gras.

En ce qui concerne la stratégie de fusion précoce, là encore, les modèles entraînés avec la carte *Tmax* ou le volume *pente maximale* sont plus performants que le modèle de référence, sauf pour le GS 3+4 (classe avec le plus grand nombre de lésions dans l'ensemble de données, voir tableau 5.6), où les sensibilités de tous les modèles sont très proches, et pour le GS 3+3 pour *Tmax*. Le modèle *wash-out* révèle également de bonnes performances et surpasse le modèle de base pour chaque GS sauf le GS 3+4 à 1.5 FP.

Modèle	GS $\geq$ 8	GS 4+3	GS 3+4	GS 3+3
Fusion précoce				
référence : IRM-bp	0.57 ± 0.06	0.42 ± 0.07	$\mathbf{0.45 \pm 0.02}$	0.14 ± 0.05
IRM-bp + pente max.	$\mathbf{0.64 \pm 0.03}$	0.46 ± 0.03	0.43 ± 0.05	$\mathbf{0.17 \pm 0.04}$
IRM-bp + % rehau.	0.62 ± 0.03	0.42 ± 0.08	0.43 ± 0.02	0.06 ± 0.04
IRM-bp + Tmax	0.61 ± 0.03	$\mathbf{0.47 \pm 0.02}$	0.44 ± 0.03	0.09 ± 0.03
IRM-bp + wash-in.	0.54 ± 0.09	0.45 ± 0.02	0.44 ± 0.05	0.10 ± 0.02
IRM-bp + wash-out	0.60 ± 0.05	0.46 ± 0.02	0.44 ± 0.02	0.16 ± 0.02
IRM-bp + Tmax + pente max.	0.58 ± 0.05	$\mathbf{0.50 \pm 0.04}$	0.41 ± 0.04	0.15 ± 0.04
Fusion intermédiaire				
référence : IRM-bp	0.56 ± 0.04	0.45 ± 0.10	0.46 ± 0.02	$\mathbf{0.20 \pm 0.02}$
IRM-bp + pente max.	$\mathbf{0.58 \pm 0.05}$	0.37 ± 0.03	0.47 ± 0.02	0.17 ± 0.03
IRM-bp + Tmax	$\mathbf{0.58 \pm 0.05}$	$\mathbf{0.51 \pm 0.03}$	$\mathbf{0.50 \pm 0.03}$	0.11 ± 0.02

TABLEAU 9.3 – Sensibilité de détection moyenne pour chaque groupe de GS à 1.5 FP. Les résultats correspondent à la moyenne des métriques obtenues sur 4 réplicats de la validation croisée à 5 plis. Les meilleurs résultats pour chaque métrique et stratégie de fusion sont en gras.

9.3.2 Prédictions visuelles

La figure 9.7 montre les prédictions des différents modèles en fusion précoce au regard de la référence IRM-bp et de la vérité terrain. Pour chacun des exemples, l'image représentée ici est issue du réplicat (parmi les 4) pour lequel le score de kappa était le plus élevé pour le pli concerné.

Le premier exemple de chacun des scanners correspond au patient d'où sont issues les cartes de perfusion de la figure 9.3, 9.5 et 9.6. Dans le premier exemple (colonne 1), le modèle biparamétrique de référence détecte la lésion GS 3+4, mais avec un Dice très faible. Par ailleurs, un FP est identifié dans la ZT. Avec l'ajout de la perfusion, ce FP n'est plus détecté. En revanche, d'autres petites lésions FP peuvent également être détectées, mais certaines d'entre elles ne seraient plus présentes après le post-traitement qui enlève les petits éléments (comme le FP détecté avec le modèle *Tmax*). Dans cet exemple, les modèles présentant le meilleur recouvrement sont le *Tmax* et les cartes de *wash-in*/*wash-out*, mais avec une erreur de GS. Dans le deuxième exemple (colonne 2), la référence montre 2 lésions en ZP, une de grade GS 3+4 et une autre plus petite de grade GS 4+3. Alors que le modèle IRM-bp de référence ne détecte pas la lésion GS 3+4 (en vert), elle est identifiée par tous les autres modèles (avec un Dice plus ou moins important) et caractérisée par le bon GS pour la plupart des modèles (excepté le *wash-in*). Concernant la deuxième lésion de grade GS 4+3, la majorité des modèles semble l'identifier (pas le *wash-in* ni le dernier modèle à deux entrées), mais avec un recouvrement assez faible et le mauvais grade (GS 3+4 prédit au lieu du grade supérieur). Le modèle entraîné avec le volume de la *pente maximale* semble mieux identifier la lésion, dont le centre coïncide davantage avec celui de la lésion prédite. Le premier exemple pour le scanner GE Discovery (colonne 3) est le seul cas où les deux lésions présentes ont été identifiées par tous les modèles et avec le bon grade (mélangé à d'autres grades pour le *% de rehaussement* et le modèle à deux entrées). Le modèle de référence montre de petits FP qui pourraient être éliminés par post-traitement. En revanche, dans ce cas, l'ajout de la carte *Tmax* détériore les résultats : une lésion FP est prédite en ZT. Cela n'est pas étonnant au vu de la carte *Tmax* présentée figure 9.5, où cette région présente un temps très court. L'ajout de la perfusion peut quelquefois mener à des FP supplémentaires.

C'est également le cas pour le premier exemple de Philips Ingenia (colonne 5) où des FP sont ajoutés avec l'inclusion de la perfusion (sauf avec le *wash-in* et le *wash-out*

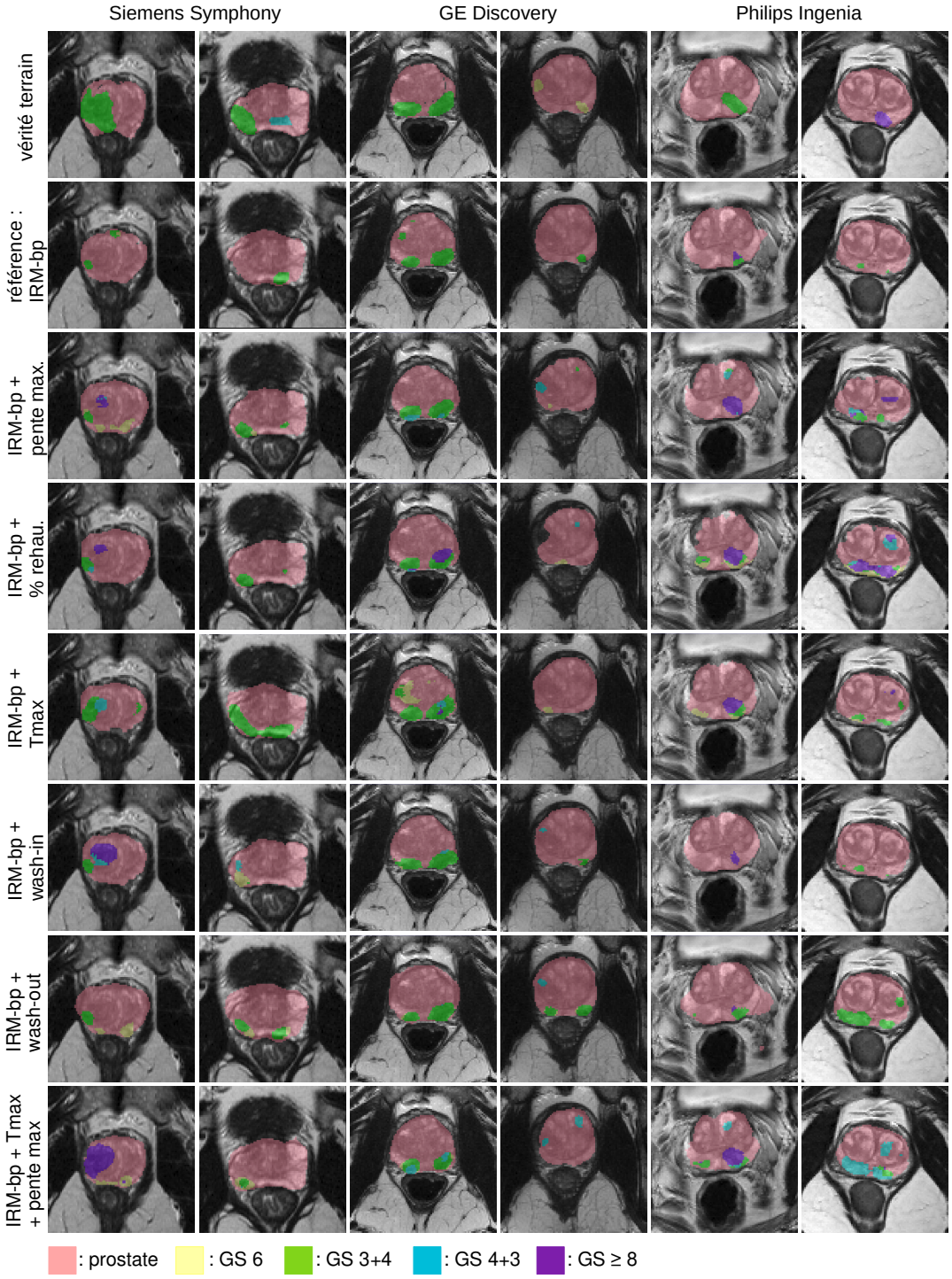


FIGURE 9.7 – Exemple de prédictions brutes (sans post-traitement) des différents modèles en fusion précoce pour des patients provenant des trois scanners. Le premier exemple de Siemens correspond à celui de la figure 9.3, le premier de GE à la figure 9.5 et le premier de Philips à la figure 9.6.

où la petite prédiction FP sera éliminée après post-traitement). Le FP prédit en ZP droite avec les modèles *% de rehaussement*, *Tmax* et *Tmax + pente maximale* a d'ailleurs été noté comme zone suspecte par les radiologues (voir la figure 9.6). Malgré tout, l'ajout de la perfusion permet d'obtenir un meilleur Dice pour cette lésion avec la

quasi-totalité des modèles.

Enfin, dans le dernier exemple (colonne 6), l'ajout de la perfusion permet avec certaines cartes d'identifier une lésion très agressive (GS $\geq$ 8) qui n'est pas repérée par le modèle biparamétrique. Une autre coupe issue de ce patient a d'ailleurs été présentée figure 7.6 et cette même lésion fait partie des cas d'échecs pour le modèle ProstAttention-Net. Ce cas de cancer mucoïde, non visible en T2-w, semble difficile, bien qu'il ait été repéré par les deux radiologues. Avec l'ajout de la perfusion, 5/6 modèles identifient la lésion, mais avec le mauvais grade (sauf pour le *% de rehaussement* qui prédit bien une lésion GS $\geq$ 8) et accompagné de FP.

Sur les exemples présentés ici, l'ajout de la carte de *wash-out* semble être la solution de premier choix, puisqu'elle permet d'identifier toutes les lésions. Toutefois, il réside des erreurs concernant l'agressivité des lésions et la majorité est prédite GS 3+4, grade le plus représenté en nombre de lésions. En outre, ces exemples ne sont qu'un échantillon des prédictions et les métriques rapportées dans les tableaux ci-dessus doivent également être considérées. Il faut également garder en tête la variabilité existante entre les 4 réplicats de validation croisée.

9.3.3 Impact des différentes cartes de perfusion selon les scanners

La figure 9.8 montre les performances de détection des lésions CS pour chacun des 4 réplicats de validation croisée, pour chaque carte de perfusion incluse, et ce, par scanner. Chaque graphique contient donc 4 courbes pour chacun des 3 scanners soit 12 courbes au total. La figure 9.8A expose les performances de référence avec le modèle biparamétrique. Une certaine variabilité est visible entre les différents réplicats mais globalement, les meilleures performances sont obtenues pour les patients issus du scanner majoritaire GE Discovery, suivi de près par les patients de Siemens Symphony (2$^{\text{ème}}$ scanner le plus représenté), puis par les Philips Ingenia qui s'avèrent être bien en dessous, comme également vu section 7.5.

L'impact des cartes de perfusion n'est pas le même pour tous les scanners :

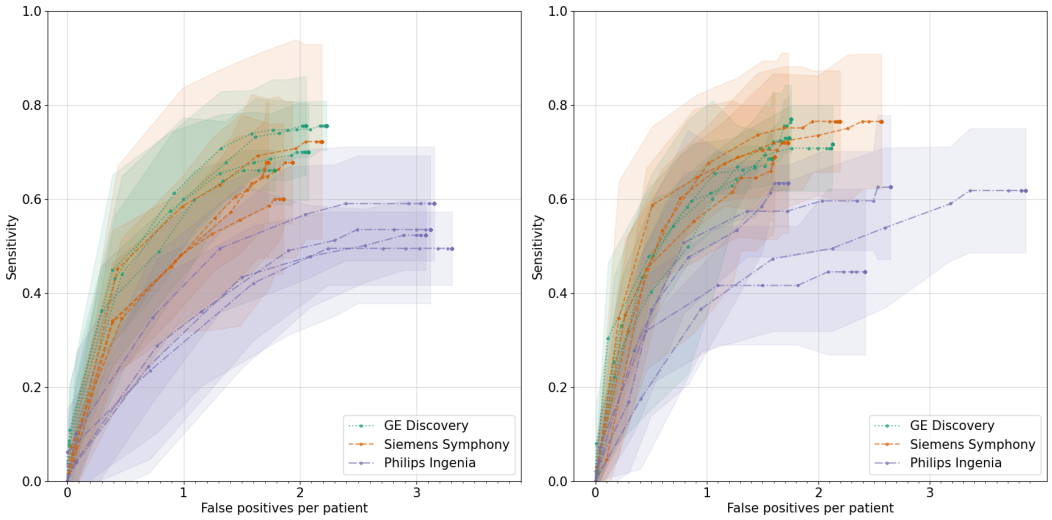
— les volumes correspondant à la *pente maximale*, les cartes de *% de rehaussement* et les cartes de *Tmax* permettent aux patients acquis sur le Siemens Symphony d'égaler les performances du GE Discovery ;

— parmi toutes les cartes, seules les cartes de *Tmax* semblent être bénéfiques aux 26 patients acquis sur le scanner Philips Ingenia, avec des performances égalant les autres scanners et un nombre de FP réduit ;

— alors que la carte *% rehaussement* semble diminuer la variabilité entre les réplicats pour les deux scanners majoritaires, elle est accrue pour les patients Philips Ingenia ;

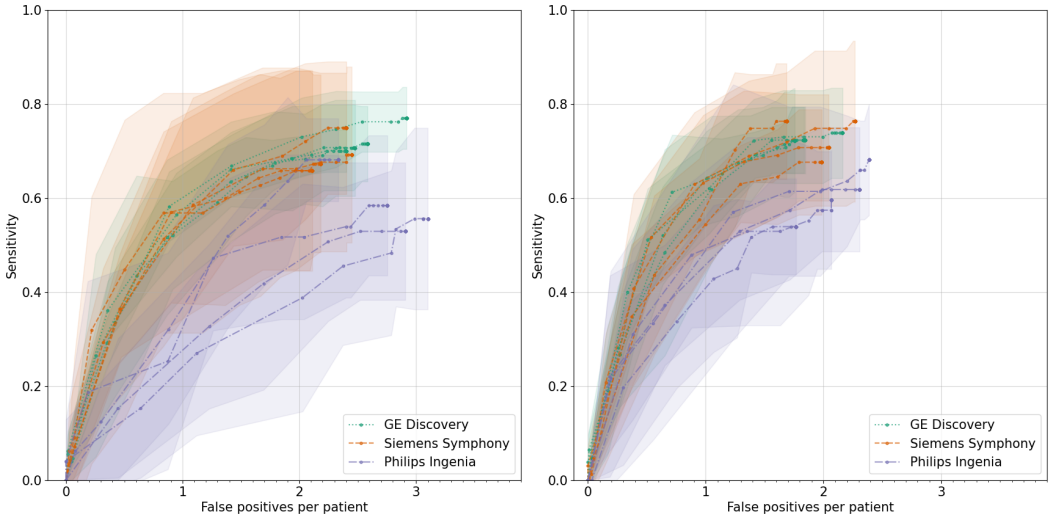
— les cartes *% rehaussement* et *wash-in* augmentent le nombre de FP pour les GE Discovery et Siemens Symphony. Il reste en revanche équivalent pour les Philips Ingenia.

À la vue de ces cartes, il semblerait que la carte de *Tmax* soit la plus bénéfique à tous les scanners, hormis pour les patients GE Discovery où les performances de référence étaient déjà élevées.



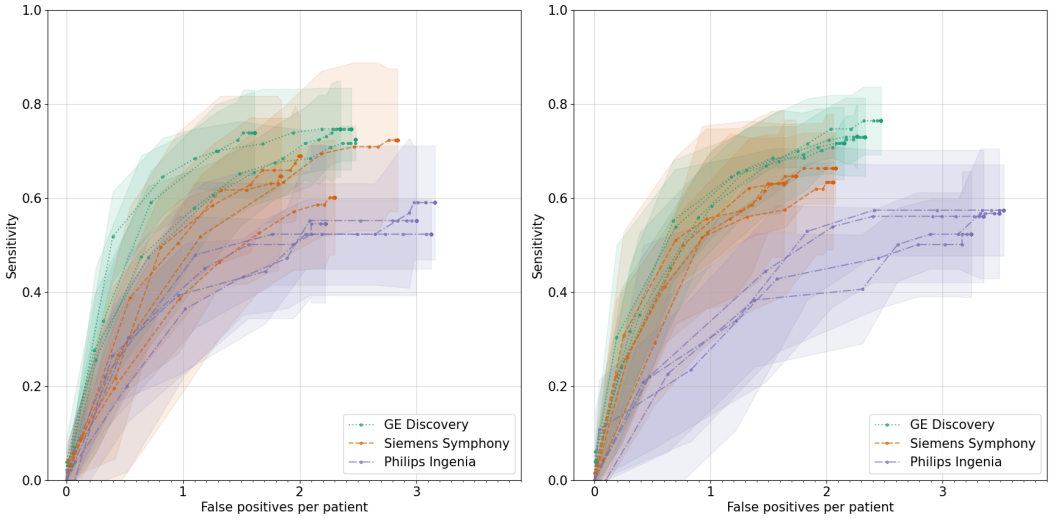
(A) référence (IRM-bp)

(B) pente maximale



(C) % rehaussement

(D) Tmax



(E) wash-in

(F) wash-out

FIGURE 9.8 – Courbes FROC obtenues par scanner pour chacun des 4 réplicats selon les différents encodages de la séquence DCE. Chaque courbe correspond à une expérience de validation croisée à 5 plis.

9.4 Discussion

Il a été observé que le gain de performance dépend des cartes de perfusion considérées et de la stratégie de fusion. Bien qu'aucune directive stricte n'ait pu être dégagée de cette étude, nous avons montré que, globalement, les volumes *Tmax*, *washout* et *pente max* amélioraient la détection des lésions de CaP. Cela est cohérent avec d'autres études où l'on a constaté que la valeur du *Tmax* était corrélée de manière significative à la présence de CaP [143].

La stratégie de fusion intermédiaire semble être bénéfique à la fois en termes de sensibilité de détection et de segmentation de la prostate.

Notre étude ne démontre pas l'impact positif des cartes *wash-in* contrairement à d'autres études radiomiques [119, 57]. Cela peut s'expliquer par la façon dont nous avons extrait les cartes de *wash-in* : la fin de la période de vitesse de rehaussement a en effet été définie au moment où l'intensité la plus haute a été observée dans le voxel, ce qui peut être atteint à la fin de l'acquisition. En outre, la vitesse de rehaussement est souvent considérée comme la pente de la droite de régression calculée à partir du signal (voir figure 2.12), et ne correspond pas directement à la pente entre le premier et le dernier temps de la période de rehaussement comme calculé ici. De plus, la normalisation de l'intensité a été effectuée après que les coupes des volumes aient été redimensionnées à une taille de 96×96 pixels, excluant ainsi la veine et l'artère iliaques externes, où un signal élevé est observé au temps d'arrivée. Cela pourrait induire une forte variabilité dans ces cartes de *wash-in*, altérant ainsi l'impact de cette carte de perfusion. Cette normalisation postérieure au redimensionnement des cartes peut également avoir un impact sur d'autres cartes de perfusion. Il est d'ailleurs possible que les cartes *Tmax* ressortent comme l'une des meilleures alternatives puisque l'exclusion de la veine et de l'artère iliaques de l'image n'impacte pas la normalisation. En effet, ces zones correspondent à des temps au pic très courts et non à des intensités très élevées, comme c'est le cas dans toutes les autres cartes. Toutes les valeurs des cartes *Tmax*, qui sont d'ailleurs dans un intervalle moins important (entre 0 et le nombre de temps d'acquisition), sont sûrement représentées dans l'image après recadrage, où l'impact d'une normalisation post recadrage sera moindre, voire nul.

Enfin, l'inclusion de cette information dynamique semble particulièrement utile pour les scanners peu représentés, dont les performances peuvent être améliorées. L'information sur chaque échantillon est ainsi plus riche, ce qui améliore particulièrement les performances des scanners dont la base d'apprentissage est de petite taille.

9.5 Conclusion et perspectives

Cette étude démontre que les cartes de perfusion dérivées des examens d'IRM DCE ont un impact positif sur les performances des modèles profonds de segmentation multiclasse du CaP. Les performances rapportées sont conformes à l'état de l'art et surpassent le kappa rapporté par De Vente et al. [27] ($\kappa = 0.172 \pm 0.169$) avec un modèle de régression ordinale basé sur l'IRM-bp.

Les perspectives seraient :

— d'étudier l'impact de la normalisation des cartes de perfusion ;

— d'inclure conjointement les trois cartes de perfusion (*Tmax*, *pente maximale*, *wash-out*) en plus des cartes T2-w et ADC dans un modèle de segmentation du CaP, avec une stratégie de fusion précoce ou intermédiaire ;

— d'analyser les distributions des paramètres en fonction des scanners. En effet, ces métriques (Tmax, etc.) ne sont pas censées dépendre du scanner, mais des caractéristiques biologiques. Il serait intéressant de regarder si cela est bien vérifié ;

— d'étudier l'influence de cette séquence dynamique sur le modèle ProstAttention-Net présenté dans le chapitre 7.

Chapitre 10

Conclusion et perspectives

10.1 Contributions

Le travail présenté dans cette thèse porte sur la mise au point d'un système CAD pour la détection du CaP à partir d'images IRM multiparamétriques. L'originalité principale réside dans la caractérisation de l'agressivité des lésions (donnée par le score de Gleason), information non décelable par les radiologues à partir des images IRM seules. Cette tâche a pu être apprise par nos modèles grâce à la base de données CLARA-P à disposition, basée sur la vérité terrain la plus fiable qu'il soit, la pièce de prostatectomie.

Dans le chapitre 7, nous avons proposé une architecture permettant d'utiliser une information anatomique apprise par le réseau, à savoir la localisation de la prostate, pour mieux détecter et caractériser les lésions par agressivité à partir d'IRM biparamétrique. Avec l'intégration de cette connaissance par l'intermédiaire d'un mécanisme d'attention, nous observons une amélioration des performances, en termes de sensibilité, mais également de précision quant au grade attribué. Ce modèle supervisé requiert des annotations complètes à la fois de la prostate et des lésions, fastidieuses à obtenir. Pour alléger ce besoin d'annotations complètes, nous avons adapté la fonction de coût afin de permettre l'inclusion de données pour lesquelles l'annotation de la prostate ou des lésions serait manquante. Nous avons montré des performances équivalentes avec 50 % des annotations de la prostate ou des lésions. Toutefois, cette approche ne permet pas l'inclusion de patients pour lesquels les contours des lésions ou de la prostate seraient partiellement connus.

Dans le chapitre 8, nous avons étudié un modèle permettant d'inclure des données partiellement annotées. L'approche évaluée exploite des annotations consistant en des points tracés dans les lésions ou la prostate grâce à une contrainte sur la taille de la prédiction associée à l'entropie croisée partielle. Les résultats obtenus s'approchent de la référence totalement supervisée et sont prometteurs pour la suite. Cette approche est particulièrement intéressante dans le contexte du CaP puisque la vérité terrain pour l'agressivité des lésions est majoritairement basée sur l'échantillon de biopsie, qui ne caractérise le GS de la lésion qu'au niveau de la biopsie et ne permet pas de connaître son étendue. La démarche faiblement supervisée permet en outre un gain de temps considérable et pourrait être envisagée pour inclure plus rapidement des données provenant d'autres bases de données, pour étoffer la base, mais aussi mieux généraliser sur cette nouvelle base. Cependant, cette étude ne se base toujours que sur l'IRM biparamétrique, alors qu'une séquence supplémentaire est acquise en pratique clinique.

Dans le chapitre 9, nous avons analysé le potentiel apport de l'imagerie dynamique ou de perfusion (DCE) dans l'entraînement d'un réseau de neurones profond

pour la segmentation des lésions par agressivité. En effet, la quasi-totalité des travaux pour la segmentation du CaP n'exploite pas cette séquence d'imagerie, mais se base sur des modèles biparamétriques. Nous avons montré que l'inclusion de la perfusion pouvait améliorer les performances des modèles, selon la manière de transformer cette information 4D en 3D. En particulier, les cartes *Tmax*, *wash-out* et le volume correspondant à la *pente maximale* semblent porter une information utile au modèle. La combinaison de deux de ces cartes (*Tmax* et *pente maximale*) en entrée ne semble pas en revanche bénéfique pour le modèle. Nous avons également étudié l'impact d'une fusion précoce ou intermédiaire au niveau de l'espace latent et montré que les performances étaient boostées par une fusion intermédiaire. Cette étude reste toutefois préliminaire et la génération des cartes de perfusion doit être étudiée davantage, notamment leur normalisation.

10.2 Limites et perspectives

Plusieurs limites et perspectives concernant les données, l'entraînement et l'évaluation des modèles peuvent être listées.

10.2.1 Données d'entraînement

Tout d'abord, plusieurs limites et perspectives concernent notre exploitation des données de la base CLARA-P.

— La vérité terrain exploitée ici est quelquefois imparfaite : il arrive qu'il y ait quelques voxels annotés comme du fond au sein de la prostate (voir colonne 3 de la figure 9.7), ou encore des lésions qui sortent de la prostate. Un prétraitement simple avec une opération de fermeture du masque binaire aurait pu être réalisé. Toutefois, les pixels concernés sont très peu nombreux et ne semblent pas impacter le processus d'apprentissage.

— La base de données CLARA-P est très riche et comporte les annotations des lésions sur les différentes séquences disponibles, pour chacun des patients. Nous avons ici considéré la vérité tracée sur la séquence T2-w, qui n'est pas forcément la modalité la plus utilisée pour prédire la présence de cancer et son grade. Une perspective serait d'exploiter les autres contours présents (notamment ceux du DWI), soit à la place des contours actuels, soit conjointement, avec un modèle qui aurait plusieurs branches de sortie fusionnées par la suite.

— La base CLARA-P contient les contours tracés par les deux radiologues lors de leur analyse des images IRM, sans qu'ils aient connaissance des résultats de la prostatectomie. Nous n'avons utilisé que les contours finaux tracés en consensus *a posteriori*, mais il pourrait être intéressant d'exploiter les contours des radiologues et l'incertitude associée.

— Aucun recalage n'a été réalisé entre les différentes séquences, peu de mouvement ayant été observé. Toutefois, il se pourrait qu'il soit malgré tout bénéfique pour certains patients d'aligner les séquences, point qu'il serait judicieux d'étudier.

— La séquence de diffusion n'a pas été incluse directement (seulement par l'intermédiaire des cartes ADC), notamment à cause des différences en termes de valeurs de b disponibles et des valeurs maximales très faibles pour certains patients (800 pour les Siemens Symphony). Une étude récente [132] a proposé une

méthode permettant d'extrapoler la diffusion pour avoir l'équivalent d'une séquence $b2000$ et enlever ainsi la variabilité due à la valeur de b, qu'il serait pertinent d'évaluer.

— Enfin, la base est très hétérogène pour ce qui est des scanners la composant et nous avons d'ailleurs observé des différences de performances selon les scanners. Il est difficile de distinguer la part due à la représentation inégale de chacun des scanners et celle due à une variabilité inter-scanner. Toutefois, des méthodes d'adaptation de domaine, à appliquer sur les images d'entrée telles que ComBat [69, 48], ou bien lors de l'apprentissage avec une branche adversaire par exemple, pourraient contrebalancer cette variabilité.

10.2.2 Entraînement des réseaux de neurones

— Une certaine variabilité a été observée entre les différentes expériences de validation croisée. Nous avons recouru à des réplicats pour y pallier, mais cela s'accompagne d'une consommation importante de ressources et allonge considérablement le temps nécessaire pour réaliser une expérience. Cette variabilité est très sûrement due au faible nombre d'exemples disponibles comparativement à la difficulté de la tâche à établir (segmentation à 6 classes) mais il serait intéressant d'étudier des méthodes permettant d'obtenir des résultats plus robustes.

— Toujours à propos des réplicats : les ayants à disposition pour un certain nombre d'expériences, il serait intéressant de les exploiter davantage, par exemple dans un modèle ensembliste, pour obtenir une mesure de certitude associée aux prédictions.

— Le critère d'arrêt considéré pour le choix du modèle à la fin de l'entraînement peut être déterminant. Dans ce travail de thèse, nous avons considéré la moyenne du Dice par classe. Cette métrique est très sévère puisqu'une lésion GS 4+3 prédite GS 3+4 donnera un Dice de 0. De plus, elle ne correspond pas à une métrique utilisée directement dans l'évaluation, bien qu'elle soit corrélée aux FROC par classe et au kappa. Exploiter directement une métrique utilisée en évaluation telle que le kappa ou la sensibilité à un taux de FP fixé pourrait permettre de sélectionner un modèle avec de meilleures performances en validation. Toutefois, il y a le risque que ce modèle généralise ensuite moins bien sur de nouvelles données.

— Les classes de nos modèles de segmentation présentent une certaine hiérarchie : une lésion GS ≥ 8 est plus maligne qu'une lésion GS $\geq 4+3$ qui l'est plus que du GS $\geq 3+4$ et ainsi de suite, jusqu'au tissu sain. Cette hiérarchie pourrait être exploitée lors de l'apprentissage, soit par l'utilisation d'une fonction de coût adaptée, telle que la kappa loss, soit par un encodage des classes ordinal, comme le font Cao et al. [15] et Abraham et al. [2], soit en réalisant une tâche de régression ordinale comme De Vente et al. [27].

10.2.3 Évaluation des modèles

— Le choix de la bonne métrique est très difficile pour la tâche évaluée dans cette thèse, qui concerne à la fois un problème de détection et de segmentation multiclasse. Ce sujet est actuellement activement discuté dans la communauté [101]. D'un côté, les courbes FROC par classe sont très pessimistes, et de l'autre le score de kappa s'avère peu robuste et ne considère que les lésions VP. Une

métrique qui pondérerait l'erreur, comme le score kappa, mais qui prendrait également en compte la sensibilité serait de grand intérêt. En outre, les métriques ne vont pas toujours dans le même sens, rendant la conclusion finale difficile.

— Parmi les métriques utilisées dans cette thèse, les courbes FROC binaires pour évaluer la détection des lésions CS ont été très utilisées. Or, ces métriques ne rapportent pas la capacité du modèle à détecter les lésions GS 6, alors qu'il s'agit d'un grade crucial : ce sont les lésions les moins détectées par les radiologues (et par les modèles), pour lesquelles la surveillance active est particulièrement intéressante.

— Il serait également intéressant de travailler sur les probabilités prédites en sortie par les modèles. En effet, nous avons toujours considéré la fonction arg max pour obtenir la classe prédite. Or, il se pourrait qu'un seuil plus faible / variable selon les classes donne de meilleurs résultats que l'argmax. Par exemple, Schelb et al. [111] fixent des seuils à 0.22 et 0.33 dans un contexte binaire pour prédire la présence de CaP (et non à 0.5 comme l'aurait fait l'argmax).

— Enfin, l'utilisation de ces modèles en condition clinique n'a pas été évaluée. Cela devrait être fait dans le cadre du projet PERFUSE, en particulier dans le cadre de l'étude multicentrique CHANGE. Les modèles devront sûrement être adaptés pour tenir compte de l'hétérogénéité de la base de données, en cours de construction. Le but final sera de fournir aux radiologues les prédictions des modèles et d'évaluer l'apport du modèle pour le diagnostic.

Contributions de l'auteur

Revues internationales avec comité de lecture

Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin et Carole Lartizien. « ProstAttention-Net : A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans ». en. In : *Medical Image Analysis* 77 (avr. 2022), p. 102347. ISSN : 1361-8415. DOI : 10.1016/j.media.2021.102347 (cf. chapitre 7)

Olivier Rouvière, Rémi Souchon, Carole Lartizien, Adeline Mansuy, Laurent Magaud, Matthieu Colom, Marine Dubreuil-Chambardel, Sabine Debeer, Tristan Jaouen, Audrey Duran, Pascal Rippert, Benjamin Riche, Caterina Monini, Virginie Vlaeminck-Guillem, Julie Haesebaert, Muriel Rabilloud et Sébastien Crouzet. « Detection of ISUP $\geq$ 2 prostate cancers using multiparametric MRI : prospective multicentre assessment of the non-inferiority of an artificial intelligence system as compared to the PI-RADS V.2.1 score (CHANGE study) ». BMJ Open 12.2 (2022). DOI : 10.1136/bmjopen-2021-051274

Tristan Jaouen, Rémi Souchon, Paul C Moldovan, Flavie Bratan, Audrey Duran, Au Hoang Dinh, Florian Di Franco, Sabine Debeer, Marine Dubreuil-Chambardel, Alain Ruffion, Nicolas Arfi, Marc Colombel, Sébastien Crouzet, Christelle Gonindard-Melodelima et Olivier Rouvière. « Characterization of high-grade prostate cancer at multiparametric MRI using a radiomic-based computer-aided diagnosis system as standalone and second reader » [en révision à *Scientific Reports*]

Congrès internationaux

Audrey Duran, Pierre-Marc Jodoin et Carole Lartizien. « Prostate Cancer Semantic Segmentation by Gleason Score Group in bi-parametric MRI with Self Attention Model on the Peripheral Zone ». en. Medical Imaging with Deep Learning. ISSN : 2640-3498. Sept. 2020 (cf. chapitre 7)

Audrey Duran, Gaspard Dussert et Carole Lartizien. « Learning to segment prostate cancer by aggressiveness from scribbles in bi-parametric MRI ». SPIE Medical Imaging 2022 : Image Processing [oral] (cf. chapitre 8)

Audrey Duran, Gaspard Dussert et Carole Lartizien. « Perfusion imaging in deep prostate cancer detection from mp-MRI : can we take advantage of it ? » 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). 2022 (cf. chapitre 9)

Congrès nationaux/régionaux

Audrey Duran, Pierre-Marc Jodoin et Carole Lartizien. « Intelligence artificielle pour la détection et caractérisation des lésions par agressivité dans l'IRM de prostate ». Forum de la recherche en cancérologie Auvergne-Rhône-Alpes. Avr. 2021 [prix du meilleur e-poster]

Audrey Duran, Pierre-Marc Jodoin et Carole Lartizien. « Segmentation des lésions cancéreuses par aggressivité dans l'IRM de prostate ». 5th meeting of Société Française de Résonance Magnétique en Biologie et Médecine, Lyon. Sept. 2021 [oral]

Bibliographie

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu et Xiaoqiang Zheng. *TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems*. 2016. arXiv : 1603.04467 [cs.DC] (cf. p. 83).

[2] Bejoy Abraham et Madhu S. Nair. « Automated grading of prostate cancer using convolutional neural network and ordinal class classifier ». en. In : *Informatics in Medicine Unlocked* 17 (jan. 2019), p. 100256. ISSN : 2352-9148. DOI : 10.1016/j.imu.2019.100256 (cf. p. 46, 52, 73, 85, 147).

[3] Bejoy Abraham et Madhu S. Nair. « Computer-aided classification of prostate cancer grade groups from MRI images using texture features and stacked sparse autoencoder ». In : *Computerized Medical Imaging and Graphics* 69 (nov. 2018), p. 60-68. ISSN : 0895-6111. DOI : 10.1016/j.compmedimag.2018.08.006 (cf. p. 46, 52, 73).

[4] Nader Aldoj, Steffen Lukas, Marc Dewey et Tobias Penzkofer. « Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network ». eng. In : *Eur Radiol* 30.2 (fév. 2020), p. 1243-1253. ISSN : 1432-1084. DOI : 10.1007/s00330-019-06417-z (cf. p. 131).

[5] Rahaf Aljundi, Jérôme Lehaire, Fabrice Prost-Boucle, Olivier Rouvière et Carole Lartizien. « Transfer Learning for Prostate Cancer Mapping Based on Multicentric MR Imaging Databases ». en. In : *Machine Learning Meets Medical Imaging*. Sous la dir. de Kanwal Bhatia et Herve Lombaert. Lecture Notes in Computer Science. Cham : Springer International Publishing, 2015, p. 74-82. ISBN : 978-3-319-27929-9. DOI : 10.1007/978-3-319-27929-9_8 (cf. p. 44).

[6] Ruba Alkadi, Fatma Taher, Ayman El-baz et Naoufel Werghi. « A Deep Learning-Based Approach for the Detection and Localization of Prostate Cancer in T2 Magnetic Resonance Images ». en. In : *Journal of Digital Imaging* (nov. 2018). DOI : 10.1007/s10278-018-0160-1 (cf. p. 44, 45, 46, 51).

[7] Michelle Bardis, Roozbeh Houshyar, Chanon Chantaduly, Karen Tran-Harding, Alexander Ushinsky, Chantal Chahine, Mark Rupasinghe, Daniel Chow et Peter Chang. « Segmentation of the Prostate Transition Zone and Peripheral Zone on MR Images with Deep Learning ». In : *Radiology : Imaging Cancer* 3.3 (mai 2021). Publisher : Radiological Society of North America, e200024. DOI : 10.1148/rycan.2021200024 (cf. p. 50, 70).

[8] Christopher M. Bishop. « Pattern Recognition and Machine Learning ». In : *springer* (2006) (cf. p. 21).

[9] Valentina Brancato, Giuseppe Di Costanzo, Luca Basso, Liberatore Tramontano, Marta Puglia, Alfonso Ragozzino et Carlo Cavaliere. « Assessment of DCE Utility for PCa Diagnosis Using PI-RADS v2.1 : Effects on Diagnostic Accuracy and Reproducibility ». en. In : *Diagnostics* 10.3 (mars 2020). Number : 3 Publisher : Multidisciplinary Digital Publishing Institute, p. 164. DOI : 10.3390/diagnostics10030164 (cf. p. 131).

[10] Flavie Bratan, Emilie Niaf, Christelle Melodelima, Anne Laure Chesnais, Rémi Souchon, Florence Mège-Lechevallier, Marc Colombel et Olivier Rouvière. « Influence of imaging and histological factors on prostate cancer detection and localisation on multiparametric MRI : a prospective study ». en. In : *Eur Radiol* 23.7 (juill. 2013), p. 2019-2029. ISSN : 1432-1084. DOI : 10.1007/s00330-013-2795-0 (cf. p. 56, 89).

[11] S. Brunelle, C. Zemmour, F. Bratan, F. Mège-Lechevallier, A. Ruffion, M. Colombel, S. Crouzet, A. Sarran et O. Rouvière. « Variability induced by the MR imager in dynamic contrast-enhanced imaging of the prostate ». en. In : *Diagnostic and Interventional Imaging* 99.4 (avr. 2018), p. 255-264. ISSN : 2211-5684. DOI : 10.1016/j.diii.2017.12.003 (cf. p. 135).

[12] Philip C. Bunch, John F. Hamilton, Gary K. Sanderson et Arthur H. Simmons. « A Free Response Approach To The Measurement And Characterization Of Radiographic Observer Performance ». In : *Application of Optical Instrumentation in Medicine VI*. T. 0127. International Society for Optics et Photonics, déc. 1977, p. 124-135. DOI : 10.1117/12.955926 (cf. p. 39).

[13] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra et Thomas J. Fuchs. « Clinical-grade computational pathology using weakly supervised deep learning on whole slide images ». en. In : *Nature Medicine* 25.8 (août 2019). Number : 8 Publisher : Nature Publishing Group, p. 1301-1309. ISSN : 1546-170X. DOI : 10.1038/s41591-019-0508-1 (cf. p. 111).

[14] Yigit B. Can, Krishna Chaitanya, Basil Mustafa, Lisa M. Koch, Ender Konukoglu et Christian F. Baumgartner. « Learning to Segment Medical Images with Scribble-Supervision Alone ». In : *arXiv :1807.04668 [cs]* (juill. 2018). arXiv : 1807.04668 (cf. p. xix, 112, 113).

[15] Ruiming Cao, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman et Kyunghyun Sung. « Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet ». en. In : *IEEE TMI* (avr. 2019). DOI : 10.1109/TMI.2019.2901928 (cf. p. xvii, xviii, 45, 47, 48, 53, 73, 76, 84, 85, 89, 102, 103, 104, 117, 120, 147).

[16] A. Chaddad, M. J. Kucharczyk, C. Desrosiers, I. P. Okuwobi, Y. Katib, M. Zhang, S. Rathore, P. Sargos et T. Niazi. « Deep Radiomic Analysis to Predict Gleason Score in Prostate Cancer ». In : *IEEE Access* 8 (2020). Conference Name : IEEE Access, p. 167767-167778. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.3023902 (cf. p. 46, 52, 73).

[17] Ahmad Chaddad, Michael J. Kucharczyk, Abbas Cheddad, Sharon E. Clarke, Lama Hassan, Shuxue Ding, Saima Rathore, Mingli Zhang, Yousef Katib, Boris Bahoric, Gad Abikhzer, Stephan Probst et Tamim Niazi. « Magnetic resonance imaging based radiomic models of prostate cancer : A narrative review ». en-GB. In : *Cancers* 13.3 (2021). Number : 3, p. 1-22. ISSN : 20726694. DOI : 10.3390/cancers13030552 (cf. p. 43).

[18] Dev P. Chakraborty. « Recent developments in imaging system assessment methodology, FROC analysis and the search model ». In : *Nucl Instrum Methods Phys Res A* 648 Supplement 1 (août 2011), S297-S301. ISSN : 0168-9002. DOI : 10.1016/j.nima.2010.11.042 (cf. p. xvii, 39).

[19] Lyndon Chan, Mahdi S. Hosseini et Konstantinos N. Plataniotis. « A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains ». en. In : *Int J Comput Vis* 129.2 (fév. 2021), p. 361-384. ISSN : 0920-5691, 1573-1405. DOI : 10.1007/s11263-020-01373-4 (cf. p. 108, 111).

[20] Sneha Chaudhari, Varun Mithal, Gungor Polatkan et Rohan Ramanath. « An Attentive Survey of Attention Models ». en. In : *arXiv :1904.02874 [cs, stat]* (déc. 2020). arXiv : 1904.02874 (cf. p. 80).

[21] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff et Hartwig Adam. « Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation ». In : *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (cf. p. 89).

[22] Yizheng Chen, Lei Xing, Lequan Yu, Hilary P. Bagshaw, Mark K. Buyyounouski et Bin Han. « Automatic intraprostatic lesion segmentation in multiparametric magnetic resonance images with proposed multiple branch UNet ». en. In : *Medical Physics* n/a.n/a (2020). _eprint : https ://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14517. ISSN : 2473-4209. DOI : https://doi.org/10.1002/mp.14517 (cf. p. xvii, 44, 45, 46, 51).

[23] Veronika Cheplygina, Marleen de Bruijne et Josien P. W. Pluim. « Not-so-supervised : A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis ». en. In : *Medical Image Analysis* 54 (mai 2019), p. 280-296. ISSN : 1361-8415. DOI : 10.1016/j.media.2019.03.009 (cf. p. 108).

[24] François Chollet et al. *Keras*. https://keras.io. 2015 (cf. p. 83).

[25] Renato Cuocolo, Albert Comelli, Alessandro Stefano, Viviana Benfante, Navdeep Dahiya, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, Anthony Yezzi et Massimo Imbriaco. « Deep Learning Whole-Gland and Zonal Prostate Segmentation on a Public MRI Dataset ». en. In : *Journal of Magnetic Resonance Imaging* 54.2 (2021). _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.27585, p. 452-459. ISSN : 1522-2586. DOI : 10.1002/jmri.27585 (cf. p. 50).

[26] Zhenzhen Dai, Eric Carver, Chang Liu, Joon Lee, Aharon Feldman, Weiwei Zong, Milan Pantelic, Mohamed Elshaikh et Ning Wen. « Segmentation of the Prostatic Gland and the Intraprostatic Lesions on Multiparametic Magnetic Resonance Imaging Using Mask Region-Based Convolutional Neural Networks ». en. In : *Advances in Radiation Oncology* 5.3 (mai 2020), p. 473-481. ISSN : 2452-1094. DOI : 10.1016/j.adro.2020.01.005 (cf. p. 44, 46, 51).

[27] Coen De Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim et Mitko Veta. « Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-parametric MRI ». In : *IEEE Trans. Biomed. Eng.* (2021), p. 1-1. ISSN : 0018-9294, 1558-2531. DOI : 10.1109/TBME.2020.2993528 (cf. p. 47, 53, 73, 77, 81, 85, 96, 102, 103, 117, 120, 143, 147).

[28] P. De Visschere, N. Lumen, P. Ost, K. Decaestecker, E. Pattyn et G. Villeirs. « Dynamic contrast-enhanced imaging has limited added value over T2-weighted imaging and diffusion-weighted imaging when using PI-RADSv2 for diagnosis of clinically significant prostate cancer in patients with elevated PSA ». en. In : *Clinical Radiology* 72.1 (jan. 2017), p. 23-32. ISSN : 0009-9260. DOI : 10.1016/j.crad.2016.09.011 (cf. p. 131).

[29] Thomas G. Dietterich, Richard H. Lathrop et Tomás Lozano-Pérez. « Solving the multiple instance problem with axis-parallel rectangles ». en. In : *Artificial Intelligence* 89.1 (jan. 1997), p. 31-71. ISSN : 0004-3702. DOI : 10.1016/S0004-3702(96)00034-3 (cf. p. 111).

[30] Florian Dubost, Hieab Adams, Pinar Yilmaz, Gerda Bortsova, Gijs van Tulder, M. Arfan Ikram, Wiro Niessen, Meike W. Vernooij et Marleen de Bruijne. « Weakly supervised object detection with 2D and 3D regression neural networks ». en. In : *Medical Image Analysis* 65 (oct. 2020), p. 101767. ISSN : 1361-8415. DOI : 10.1016/j.media.2020.101767 (cf. p. 110, 118).

[31] John Duchi, Elad Hazan et Yoram Singer. « Adaptive subgradient methods for online learning and stochastic optimization. » In : *Journal of machine learning research* 12.7 (2011) (cf. p. 32).

[32] Audrey Duran, Gaspard Dussert et Carole Lartizien. « Learning to segment prostate cancer by aggressiveness from scribbles in bi-parametric MRI ». In : *Medical Imaging 2022 : Image Processing*. T. 12032. SPIE, 2022, p. 178-184. DOI : 10.1117/12.2607502 (cf. p. 107).

[33] Audrey Duran, Gaspard Dussert et Carole Lartizien. « Perfusion Imaging in Deep Prostate Cancer Detection from MP-MRI : Can We Take Advantage of it ? » In : *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. 2022, p. 1-5. DOI : 10.1109/ISBI52829.2022.9761616 (cf. p. 131).

[34] Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin et Carole Lartizien. « ProstAttention-Net : A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans ». en. In : *Medical Image Analysis* 77 (avr. 2022), p. 102347. ISSN : 1361-8415. DOI : 10.1016/j.media.2021.102347 (cf. p. 79).

[35] Audrey Duran, Pierre-Marc Jodoin et Carole Lartizien. « Prostate Cancer Semantic Segmentation by Gleason Score Group in bi-parametric MRI with Self Attention Model on the Peripheral Zone ». en. In : *Medical Imaging with Deep Learning*. ISSN : 2640-3498. PMLR, sept. 2020, p. 193-204 (cf. p. xix, 79, 80, 93, 94, 95, 99).

[42] James P. Egan, Gordon Z. Greenberg et Arthur I. Schulman. « Operating Characteristics, Signal Detectability, and the Method of Free Response ». en. In : *The Journal of the Acoustical Society of America* 33.8 (1961). Publisher : Acoustical Society of AmericaASA, p. 993. ISSN : 0001-4966. DOI : 10.1121/1.1908935 (cf. p. 39).

[43] Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, Peter A. Humphrey et Grading Committee. « The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma : Definition of Grading Patterns and Proposal for a New Grading System ». eng. In : *Am. J. Surg. Pathol.* 40.2 (fév. 2016), p. 244-252. ISSN : 1532-0979. DOI : 10.1097/PAS.0000000000000530 (cf. p. 8, 73).

[44] Jonathan I. Epstein, Zhaoyong Feng, Bruce J. Trock et Phillip M. Pierorazio. « Upgrading and Downgrading of Prostate Cancer from Biopsy to Radical Prostatectomy : Incidence and Predictive Factors Using the Modified Gleason Grading System and Factoring in Tertiary Grades ». In : *Eur Urol* 61.5 (mai 2012), p. 1019-1024. ISSN : 0302-2838. DOI : 10.1016/j.eururo.2012.01.050 (cf. p. 2).

[45] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau et Sebastian Thrun. « Dermatologist-level classification of skin cancer with deep neural networks ». en. In : *Nature* 542.7639 (fév. 2017). Bandiera_abtest : a Cg_type : Nature Research Journals Number : 7639 Primary_atype : Research Publisher : Nature Publishing Group Subject_term : Diagnosis ;Machine learning ;Skin cancer Subject_term_id : diagnosis ;machine-learning ;skin-cancer, p. 115-118. ISSN : 1476-4687. DOI : 10.1038/nature21056 (cf. p. 42).

[46] Baowei Fei. « Computer-aided diagnosis of prostate cancer with MRI ». In : *Current Opinion in Biomedical Engineering*. New Developments in Biomedical Imaging 3 (sept. 2017), p. 20-27. ISSN : 2468-4511. DOI : 10.1016/j.cobme.2017.09.009 (cf. p. 43).

[47] Xinyang Feng, Jie Yang, Andrew F. Laine et Elsa D. Angelini. « Discriminative Localization in CNNs for Weakly-Supervised Segmentation of Pulmonary Nodules ». en. In : *arXiv :1707.01086 [cs]* 10435 (2017). arXiv : 1707.01086, p. 568-576. DOI : 10.1007/978-3-319-66179-7_65 (cf. p. 109).

[48] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A. Elliott, Kosha Ruparel, David R. Roalf, Theodore D. Satterthwaite, Ruben C. Gur, Raquel E. Gur, Robert T. Schultz, Ragini Verma et Russell T. Shinohara. « Harmonization of multi-site diffusion tensor imaging data ». en. In : *NeuroImage* 161 (nov. 2017), p. 149-170. ISSN : 1053-8119. DOI : 10.1016/j.neuroimage.2017.08.047 (cf. p. 147).

[49] Leo Gautheron, Ievgen Redko et Carole Lartizien. « Feature Selection for Unsupervised Domain Adaptation Using Optimal Transport ». en. In : *Machine Learning and Knowledge Discovery in Databases*. Sous la dir. de Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley et Georgiana Ifrim. Lecture Notes in Computer Science. Cham : Springer International Publishing, 2019, p. 759-776. ISBN : 978-3-030-10928-8. DOI : 10.1007/978-3-030-10928-8_45 (cf. p. 44).

[50] Nooshin Ghavami, Yipeng Hu, Eli Gibson, Ester Bonmati, Mark Emberton, Caroline M. Moore et Dean C. Barratt. « Automatic segmentation of prostate MRI using convolutional neural networks : Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration ». en. In : *Medical Image Analysis* 58 (déc. 2019), p. 101558. ISSN : 1361-8415. DOI : 10.1016/j.media.2019.101558 (cf. p. 50).

[51] David Gillespie, Connah Kendrick, Ian Boon, Cheng Boon, Tim Rattay et Moi Hoon Yap. « Deep learning in magnetic resonance prostate segmentation : A review and a new perspective ». en. In : *arXiv :2011.07795 [cs, eess]* (nov. 2020). arXiv : 2011.07795 (cf. p. 42).

[52] Ian Goodfellow, Yoshua Bengio et Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016 (cf. p. 21).

[53] L. Grady. « Random Walks for Image Segmentation ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.11 (nov. 2006). Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 1768-1783. ISSN : 1939-3539. DOI : 10.1109/TPAMI.2006.233 (cf. p. 113).

[54] Matthew D. Greer, Anna M. Brown, Joanna H. Shih, Ronald M. Summers, Jamie Marko, Yan Mee Law, Sandeep Sankineni, Arvin K. George, Maria J. Merino, Peter A. Pinto, Peter L. Choyke et Baris Turkbey. « Accuracy and agreement of PIRADSv2 for prostate cancer mpMRI : A multireader study ». en. In : *Journal of Magnetic Resonance Imaging* 45.2 (2017). _eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.25372, p. 579-585. ISSN : 1522-2586. DOI : 10.1002/jmri.25372 (cf. p. 17).

[55] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega et Dale R. Webster. « Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs ». en. In : *JAMA* 316.22 (déc. 2016), p. 2402. ISSN : 0098-7484. DOI : 10.1001/jama.2016.17216 (cf. p. 41).

[56] Geoffrey Hinton, Nitish Srivastava et Kevin Swersky. *Neural networks for machine learning lecture 6a overview of mini-batch gradient descent*. 2012 (cf. p. 32).

[57] Au Hoang Dinh, Christelle Melodelima, Rémi Souchon, Jérôme Lehaire, Flavie Bratan, Florence Mège-Lechevallier, Alain Ruffion, Sébastien Crouzet, Marc Colombel et Olivier Rouvière. « Quantitative Analysis of Prostate Multiparametric MR Images for Detection of Aggressive Prostate Cancer in the Peripheral Zone : A Multiple Imager Study ». In : *Radiology* 280.1 (fév. 2016), p. 117-127. ISSN : 0033-8419. DOI : 10.1148/radiol.2016151406 (cf. p. 143).

[58] Jie Hu, Li Shen et Gang Sun. « Squeeze-and-Excitation Networks ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2018, p. 7132-7141 (cf. p. xvii, 80, 81).

[59] Sergey Ioffe et Christian Szegedy. « Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift ». In : *ArXiv* abs/1502.03167 (2015) (cf. p. 82).

[60] Fabian Isensee, Paul F. Jäger, Simon A. A. Kohl, Jens Petersen et Klaus H. Maier-Hein. « Automated Design of Deep Learning Methods for Biomedical Image Segmentation ». In : *Nat Methods* 18.2 (fév. 2021). arXiv : 1904.08128, p. 203-211. ISSN : 1548-7091, 1548-7105. DOI : 10.1038/s41592-020-01008-z (cf. p. 50).

[61] Junichiro Ishioka, Yoh Matsuoka, Sho Uehara, Yosuke Yasuda, Toshiki Kijima, Soichiro Yoshida, Minato Yokoyama, Kazutaka Saito, Kazunori Kihara, Noboru Numao, Tomo Kimura, Kosei Kudo, Itsuo Kumazawa et Yasuhisa Fujii. « Computer-aided diagnosis of prostate cancer on magnetic resonance

imaging using a convolutional neural network algorithm ». en. In : *BJU International* 122.3 (2018), p. 411-417. ISSN : 1464-410X. DOI : 10.1111/bju.14397 (cf. p. 44, 51, 70).

[62] Mohammadhassan Izadyyazdanabadi, Evgenii Belykh, Claudio Cavallo, Xiaochun Zhao, Sirin Gandhi, Leandro Borba Moreira, Jennifer Eschbacher, Peter Nakaji, Mark C. Preul et Yezhou Yang. « Weakly-Supervised Learning-Based Feature Localization for Confocal Laser Endomicroscopy Glioma Images ». In : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Sous la dir. d'Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López et Gabor Fichtinger. T. 11071. Series Title : Lecture Notes in Computer Science. Cham : Springer International Publishing, 2018, p. 300-308. ISBN : 978-3-030-00933-5 978-3-030-00934-2. DOI : 10.1007/978-3-030-00934-2_34 (cf. p. 109).

[64] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero et Yoshua Bengio. « The One Hundred Layers Tiramisu : Fully Convolutional DenseNets for Semantic Segmentation ». en. In : *arXiv :1611.09326 [cs]* (oct. 2017). arXiv : 1611.09326 (cf. p. 80).

[65] Zhanghexuan Ji, Yan Shen, Chunwei Ma et Mingchen Gao. « Scribble-Based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation ». en. In : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Sous la dir. de Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap et Ali Khan. Lecture Notes in Computer Science. Cham : Springer International Publishing, 2019, p. 175-183. ISBN : 978-3-030-32248-9. DOI : 10.1007/978-3-030-32248-9_20 (cf. p. 113).

[66] Z. Jia, X. Huang, E. I. Chang et Y. Xu. « Constrained Deep Weak Supervision for Histopathology Image Segmentation ». In : *IEEE Transactions on Medical Imaging* 36.11 (nov. 2017). Conference Name : IEEE Transactions on Medical Imaging, p. 2376-2388. ISSN : 1558-254X. DOI : 10.1109/TMI.2017.2724070 (cf. p. 111, 112).

[67] Changzhe Jiao, Xiaoxiao Du et Alina Zare. « Addressing the Inevitable Imprecision : Multiple Instance Learning for Hyperspectral Image Analysis ». In : *Hyperspectral Image Analysis : Advances in Machine Learning and Signal Processing*. Sous la dir. de Saurabh Prasad et Jocelyn Chanussot. Cham : Springer International Publishing, 2020, p. 141-185. ISBN : 978-3-030-38617-7. DOI : 10.1007/978-3-030-38617-7_6 (cf. p. xix, 111).

[68] Yao Jin, Guang Yang, Ying Fang, Ruipeng Li, Xiaomei Xu, Yongkai Liu et Xiaobo Lai. « 3D PBV-Net : An automated prostate MRI data segmentation method ». en. In : *Computers in Biology and Medicine* 128 (jan. 2021), p. 104160. ISSN : 0010-4825. DOI : 10.1016/j.compbiomed.2020.104160 (cf. p. 42, 50).

[69] W. Evan Johnson, Cheng Li et Ariel Rabinovic. « Adjusting batch effects in microarray expression data using empirical Bayes methods ». eng. In : *Biostatistics* 8.1 (jan. 2007), p. 118-127. ISSN : 1465-4644. DOI : 10.1093/biostatistics/kxj037 (cf. p. 147).

[70] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov et Ismail Ben Ayed. « Constrained-CNN losses for weakly supervised segmentation ». en. In : *Medical Image Analysis* 54 (mai 2019), p. 88-99. ISSN : 1361-8415 (cf. p. xii, xix, 113, 118, 120, 122, 127, 129).

[71] Zia Khan, Norashikin Yahya, Khaled Alsaih, Mohammed Isam Al-Hiyali et Fabrice Meriaudeau. « Recent Automatic Segmentation Algorithms of MRI Prostate Regions : A Review ». In : *IEEE Access* 9 (2021). Conference Name : IEEE Access, p. 97878-97905. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2021.3090825 (cf. p. 42).

[72] Diederik P. Kingma et Jimmy Ba. « Adam : A Method for Stochastic Optimization ». In : *arXiv e-prints* (2014), arXiv :1412.6980 (cf. p. 32, 83).

[73] J. R. Landis et G. G. Koch. « The measurement of observer agreement for categorical data ». eng. In : *Biometrics* 33.1 (mars 1977), p. 159-174. ISSN : 0006-341X (cf. p. xxiii, 38).

[74] *Le cancer de la prostate*. https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers/Les-cancers-les-plus-frequents/Cancer-de-la-prostate. [Accédé le 23 juillet 2021] (cf. p. 6).

[75] Dong-Hyun Lee et al. « Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks ». In : *Workshop on challenges in representation learning, ICML*. T. 3. 2. 2013, p. 896 (cf. p. 114).

[76] Hyeonsoo Lee et Won-Ki Jeong. « Scribble2Label : Scribble-Supervised Cell Segmentation via Self-generating Pseudo-Labels with Consistency ». en. In : *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Sous la dir. d'Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu et Leo Joskowicz. Lecture Notes in Computer Science. Cham : Springer International Publishing, 2020, p. 14-23. ISBN : 978-3-030-59710-8. DOI : 10.1007/978-3-030-59710-8_2 (cf. p. 114).

[77] Jerome Lehaire, Rémi Flamary, Olivier Rouvière et Carole Lartizien. « Computer-aided diagnostic system for prostate cancer detection and characterization combining learned dictionaries and supervised classification ». In : *2014 IEEE International Conference on Image Processing (ICIP)*. ISSN : 2381-8549. Oct. 2014, p. 2251-2255. DOI : 10.1109/ICIP.2014.7025456 (cf. p. 44).

[78] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C. Vilanova, Paul M. Walker et Fabrice Meriaudeau. « Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI : A review ». In : *Computers in Biology and Medicine* 60 (mai 2015), p. 8-31. ISSN : 0010-4825. DOI : 10.1016/j.compbiomed.2015.02.009 (cf. p. 43, 73).

[79] Yin Li, Jian Sun, Chi-Keung Tang et Heung-Yeung Shum. « Lazy snapping ». In : *ACM Trans. Graph.* 23.3 (août 2004), p. 303-308. ISSN : 0730-0301. DOI : 10.1145/1015706.1015719 (cf. p. 113).

[80] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He et Jian Sun. « ScribbleSup : Scribble-Supervised Convolutional Networks for Semantic Segmentation ». In : 2016, p. 3159-3167 (cf. p. 114).

[81] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He et Piotr Dollar. « Focal Loss for Dense Object Detection ». In : 2017, p. 2980-2988 (cf. p. 102).

[82] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer et Henkjan Huisman. *ProstateX Challenge data*. The Cancer Imaging Archive. 2017 (cf. p. 46, 73, 96, 117, 131).

[83] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip "Eddie" Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman et Anant Madabhushi. « Evaluation of prostate segmentation algorithms for MRI : The PROMISE12 challenge ». In : *Medical Image Analysis* 18.2 (fév. 2014), p. 359-373. ISSN : 1361-8415. DOI : 10.1016/j.media.2013.12.002 (cf. p. 42).

[84] Lizhi Liu, Zhiqiang Tian, Zhenfeng Zhang et Baowei Fei. « Computer-aided Detection of Prostate Cancer with MRI : Technology and Applications ». en. In : *Academic Radiology* 23.8 (août 2016), p. 1024-1046. ISSN : 10766332. DOI : 10.1016/j.acra.2016.03.010 (cf. p. 43).

[85] Quande Liu, Qi Dou, Lequan Yu et Pheng Ann Heng. « MS-Net : Multi-Site Network for Improving Prostate Segmentation With Heterogeneous MRI Data ». In : *IEEE Transactions on Medical Imaging* 39.9 (sept. 2020). Conference Name : IEEE Transactions on Medical Imaging, p. 2713-2724. ISSN : 1558-254X. DOI : 10.1109/TMI.2020.2974574 (cf. p. 43, 50, 70).

[86] Y. Liu, G. Yang, M. Hosseiny, A. Azadikhah, S. A. Mirak, Q. Miao, S. S. Raman et K. Sung. « Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation ». In : *IEEE Access* 8 (2020). Conference Name : IEEE Access, p. 151817-151828. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2020.3017168 (cf. p. 50).

[87] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang et Anne L. Martel. « Loss odyssey in medical image segmentation ». en. In : *Medical Image Analysis* 71 (juill. 2021), p. 102035. ISSN : 1361-8415. DOI : 10.1016/j.media.2021.102035 (cf. p. 82).

[88] Andrew L. Maas, Awni Y. Hannun et Andrew Y. Ng. « Rectifier Nonlinearities Improve Neural Network Acoustic Models ». In : *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. 2013 (cf. p. 82).

[89] Anant Madabhushi et Michael Feldman. *Fused Radiology-Pathology Prostate Dataset*. Type : dataset. 2016. DOI : 10.7937/K9/TCIA.2016.TLPMR1AM (cf. p. 73).

[90] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang et Stephen Paul Smolley. « Least Squares Generative Adversarial Networks ». In : 2017, p. 2794-2802 (cf. p. 115).

[91] Charles E Metz. « ROC methodology in radiologic imaging. » In : *Investigative radiology* 21.9 (1986), p. 720-733 (cf. p. 37).

[92] Tom M Mitchell et al. « Machine learning. 1997 ». In : *Burr Ridge, IL : McGraw Hill* 45.37 (1997), p. 870-877 (cf. p. 22).

[93] Nicolas Mottet, Roderick C. N. van den Bergh, Erik Briers, Thomas Van den Broeck, Marcus G. Cumberbatch, Maria De Santis, Stefano Fanti, Nicola Fossati, Giorgio Gandaglia, Silke Gillessen, Nikos Grivas, Jeremy Grummet, Ann M. Henry, Theodorus H. van der Kwast, Thomas B. Lam, Michael Lardas, Matthew Liew, Malcolm D. Mason, Lisa Moris, Daniela E. Oprea-Lager, Henk G. van der Poel, Olivier Rouvière, Ivo G. Schoots, Derya Tilki, Thomas Wiegel, Peter-Paul M. Willemse et Philip Cornford. « EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer—2020 Update. Part 1 : Screening,

Diagnosis, and Local Treatment with Curative Intent ». en. In : *European Urology* 79.2 (fév. 2021), p. 243-262. ISSN : 0302-2838. DOI : 10.1016/j.eururo.2020.09.042 (cf. p. 9).

[94] Emilie Niaf. « Aide au diagnostic du cancer de la prostate par IRM multiparamétrique : une approche par classification supervisée ». fr. Thèse de doct. Université Claude Bernard - Lyon I, déc. 2012 (cf. p. xv, xvii, 14, 15, 16, 59).

[95] Emilie Niaf, Carole Lartizien, Flavie Bratan, Laurent Roche, Muriel Rabilloud, Florence Mège-Lechevallier et Olivier Rouvière. « Prostate Focal Peripheral Zone Lesions : Characterization at Multiparametric MR Imaging—Influence of a Computer-aided Diagnosis System ». In : *Radiology* 271.3 (mars 2014), p. 761-769. ISSN : 0033-8419. DOI : 10.1148/radiol.14130448 (cf. p. 43).

[96] Emilie Niaf, Olivier Rouvière, Florence Mège-Lechevallier, Flavie Bratan et Carole Lartizien. « Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI ». en. In : *Phys. Med. Biol.* 57.12 (mai 2012). Publisher : IOP Publishing, p. 3833-3851. ISSN : 0031-9155. DOI : 10.1088/0031-9155/57/12/3833 (cf. p. 43).

[97] Aytekin Oto, Arda Kayhan, Yulei Jiang, Maria Tretiakova, Cheng Yang, Tatjana Antic, Farid Dahi, Arieh L. Shalhav, Gregory Karczmar et Walter M. Stadler. « Prostate cancer : differentiation of central gland cancer from benign prostatic hyperplasia by using diffusion-weighted and dynamic contrast-enhanced MR imaging ». eng. In : *Radiology* 257.3 (déc. 2010), p. 715-723. ISSN : 1527-1315. DOI : 10.1148/radiol.10100021 (cf. p. 86).

[98] Adam Paszke, Abhishek Chaurasia, Sangpil Kim et Eugenio Culurciello. « ENet : A Deep Neural Network Architecture for Real-Time Semantic Segmentation ». en. In : *arXiv :1606.02147 [cs]* (juin 2016). arXiv : 1606.02147 (cf. p. 89).

[99] J. Peng et Y. Wang. « Medical Image Segmentation With Limited Supervision : A Review of Deep Network Models ». In : *IEEE Access* 9 (2021). Conference Name : IEEE Access, p. 36827-36851. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2021.3062380 (cf. p. 108).

[100] P. Puech, A. Sufana Iancu, B. Renard, A. Villers et L. Lemaitre. « Detecting prostate cancer with MRI — why and how ». en. In : *Diagnostic and Interventional Imaging*. Urologic oncology 93.4 (avr. 2012), p. 268-278. ISSN : 2211-5684. DOI : 10.1016/j.diii.2012.01.019 (cf. p. xvi, 18).

[101] Annika Reinke, Matthias Eisenmann, Minu D. Tizabi, Carole H. Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M. Jorge Cardoso, Veronika Cheplygina, Keyvan Farahani, Ben Glocker, Doreen Heckmann-Nötzel, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Jens Kleesiek, Tahsin Kurc, Michal Kozubek, Bennett A. Landman, Geert Litjens, Klaus Maier-Hein, Bjoern Menze, Henning Müller, Jens Petersen, Mauricio Reyes, Nicola Rieke, Bram Stieltjes, Ronald M. Summers, Sotirios A. Tsaftaris, Bram van Ginneken, Annette Kopp-Schneider, Paul Jäger et Lena Maier-Hein. « Common Limitations of Image Processing Metrics : A Picture Story ». en. In : *arXiv :2104.05642 [cs, eess]* (avr. 2021). arXiv : 2104.05642 (cf. p. 104, 147).

[102] Olaf Ronneberger, Philipp Fischer et Thomas Brox. « U-Net : Convolutional Networks for Biomedical Image Segmentation ». en. In : *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Sous la dir. de Nassir Navab, Joachim Hornegger, William M. Wells et Alejandro F. Frangi. Lecture Notes in Computer Science. Cham : Springer International Publishing, 2015, p. 234-241. ISBN : 978-3-319-24574-4. DOI : 10.1007/978-3-319-24574-4_28 (cf. p. xvi, 33, 82, 114, 118).

[103] F. Rosenblatt. « The perceptron : a probabilistic model for information storage and organization in the brain. » In : *Psychological review* (1958). DOI : 10.1037/H0042519 (cf. p. 24).

[105] A. G. Roy, N. Navab et C. Wachinger. « Recalibrating Fully Convolutional Networks With Spatial and Channel "Squeeze and Excitation" Blocks ». In : *IEEE Transactions on Medical Imaging* 38.2 (fév. 2019). Conference Name : IEEE Transactions on Medical Imaging, p. 540-549. ISSN : 1558-254X. DOI : 10.1109/TMI.2018.2867261 (cf. p. 80).

[106] F. Rozet, P. Mongiat-Artus, C. Hennequin, J. B. Beauval, P. Beuzeboc, L. Cormier, G. Fromont-Hankard, R. Mathieu, G. Ploussard, R. Renard-Penna, I. Brenot-Rossi, F. Bruyere, A. Cochet, G. Crehange, O. Cussenot, T. Lebret, X. Rebillard, M. Soulié, L. Brureau et A. Méjean. « Recommandations françaises du Comité de cancérologie de l'AFU – actualisation 2020–2022 : cancer de la prostate ». fr. In : *Progrès en Urologie* 30.12, Supplement (nov. 2020), S136-S251. ISSN : 1166-7087. DOI : 10.1016/S1166-7087(20)30752-1 (cf. p. 5, 6).

[107] David E Rumelhart, Geoffrey E Hinton et Ronald J Williams. « Learning representations by back-propagating errors ». In : *nature* 323.6088 (1986), p. 533-536 (cf. p. 26).

[108] Leonardo Rundo, Changhee Han, Yudai Nagano, Jin Zhang, Ryuichiro Hataya, Carmelo Militello, Andrea Tangherloni, Marco S. Nobile, Claudio Ferretti, Daniela Besozzi, Maria Carla Gilardi, Salvatore Vitabile, Giancarlo Mauri, Hideki Nakayama et Paolo Cazzaniga. « USE-Net : Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets ». en. In : *Neurocomputing* 365 (nov. 2019), p. 31-43. ISSN : 0925-2312. DOI : 10.1016/j.neucom.2019.07.006 (cf. p. 43, 50, 70, 80, 102).

[109] Anindo Saha, Matin Hosseinzadeh et Henkjan Huisman. « End-to-end prostate cancer detection in bpMRI via 3D CNNs : Effects of attention mechanisms, clinical priori and decoupled false positive reduction ». en. In : *Medical Image Analysis* 73 (oct. 2021), p. 102155. ISSN : 1361-8415. DOI : 10.1016/j.media.2021.102155 (cf. p. xvii, 44, 45, 46, 51, 73, 76, 81, 102).

[110] Hemamali Samaratunga, Rodolfo Montironi, Lawrence True, Jonathan I. Epstein, David F. Griffiths, Peter A. Humphrey, Theo van der Kwast, Thomas M. Wheeler, John R. Srigley, Brett Delahunt et Lars Egevad. « International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 1 : specimen handling ». en. In : *Modern Pathology* 24.1 (jan. 2011). Number : 1 Publisher : Nature Publishing Group, p. 6-15. ISSN : 1530-0285. DOI : 10.1038/modpathol.2010.178 (cf. p. 56).

[111] Patrick Schelb, Simon Kohl, Jan Philipp Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Sebastian Bickelhaupt, Tristan Anselm Kuder, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Klaus H. Maier-Hein et David Bonekamp. « Classification of Cancer at Prostate MRI : Deep Learning versus Clinical PI-RADS Assessment ». In : *Radiology* (oct. 2019), p. 190938. ISSN : 0033-8419. DOI : 10.1148/radiol.2019190938 (cf. p. 44, 46, 51, 70, 148).

[112] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker et Daniel Rueckert. « Attention gated networks : Learning to leverage salient regions in medical images ». In : *Medical Image Analysis* 53 (2019), p. 197-207 (cf. p. 80, 89).

[113] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh et Dhruv Batra. « Grad-CAM : Visual Explanations From Deep Networks via Gradient-Based Localization ». In : 2017, p. 618-626 (cf. p. 110).

[114] Julio Silva-Rodríguez, Adrián Colomer et Valery Naranjo. « WeGleNet : A weakly-supervised convolutional neural network for the semantic segmentation of Gleason grades in prostate histology images ». en. In : *Computerized Medical Imaging and Graphics* 88 (mars 2021), p. 101846. ISSN : 0895-6111. DOI : 10.1016/j.compmedimag.2020.101846 (cf. p. xix, 112).

[115] Geoffrey A. Sonn, Richard E. Fan, Pejman Ghanouni, Nancy N. Wang, James D. Brooks, Andreas M. Loening, Bruce L. Daniel, Katherine J. To'o, Alan E. Thong et John T. Leppert. « Prostate Magnetic Resonance Imaging Interpretation Varies Substantially Across Radiologists ». en. In : *European Urology Focus* 5.4 (juill. 2019), p. 592-599. ISSN : 2405-4569. DOI : 10.1016/j.euf.2017.11.010 (cf. p. 17).

[116] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin et M. Jorge Cardoso. « Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations ». In : *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2017)* 2017 (2017), p. 240-248. DOI : 10.1007/978-3-319-67558-9_28 (cf. p. 35).

[117] Yohan Sumathipala, Nathan Lay, Baris Turkbey, Clayton Smith, Peter L. Choyke et Ronald M. Summers. « Prostate cancer detection from multi-institution multiparametric MRIs using deep convolutional neural networks ». In : *JMI* 5.4 (déc. 2018), p. 044507. ISSN : 2329-4302, 2329-4310. DOI : 10.1117/1.JMI.5.4.044507 (cf. p. 44, 51).

[118] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal et Freddie Bray. « Global cancer statistics 2020 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries ». eng. In : *CA Cancer J Clin* (fév. 2021). ISSN : 1542-4863. DOI : 10.3322/caac.21660 (cf. p. 1, 5).

[119] Yu Sub Sung, Heon-Ju Kwon, Bum-Woo Park, Gyunggoo Cho, Chang Kyung Lee, Kyoung-Sik Cho et Jeong Kon Kim. « Prostate Cancer Detection on Dynamic Contrast-Enhanced MRI : Computer-Aided Diagnosis Versus Single Perfusion Parameter Maps ». In : *American Journal of Roentgenology* 197.5 (nov. 2011). Publisher : American Roentgen Ray Society, p. 1122-1129. ISSN : 0361-803X. DOI : 10.2214/AJR.10.6062 (cf. p. 143).

[120] Alexey Surov, Hans Jonas Meyer et Andreas Wienke. « Correlations between Apparent Diffusion Coefficient and Gleason Score in Prostate Cancer : A Systematic Review ». en. In : *European Urology Oncology* 3.4 (août 2020), p. 489-497. ISSN : 2588-9311. DOI : 10.1016/j.euo.2018.12.006 (cf. p. 16).

[121] John A Swets. « Form of empirical ROCs in discrimination and diagnostic tasks : implications for theory and measurement of performance. » In : *Psychological bulletin* 99.2 (1986), p. 181 (cf. p. 37).

[122] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu et Xiaowei Ding. « Embracing imperfect datasets : A review of deep learning solutions for medical image segmentation ». en. In : *Medical Image Analysis* 63 (juill. 2020), p. 101693. ISSN : 1361-8415. DOI : 10.1016/j.media.2020.101693 (cf. p. xix, 108, 109).

[123] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov et Christopher Schroers. « Normalized Cut Loss for Weakly-Supervised CNN Segmentation ». en. In : *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT : IEEE, juin 2018, p. 1818-1827. ISBN : 978-1-5386-6420-9. DOI : 10.1109/CVPR.2018.00195 (cf. p. 34).

[124] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers et Yuri Boykov. « On Regularized Losses for Weakly-supervised CNN Segmentation ». In : 2018, p. 507-522 (cf. p. 114, 115).

[125] Y. Tsehay, N. Lay, X. Wang, J. T. Kwak, B. Turkbey, P. Choyke, P. Pinto, B. Wood et R. M. Summers. « Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI ». In : *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. Avr. 2017, p. 642-645. DOI : 10.1109/ISBI.2017.7950602 (cf. p. 116, 124).

[126] Jasper J. Twilt, Kicky G. van Leeuwen, Henkjan J. Huisman, Jurgen J. Fütterer et Maarten de Rooij. « Artificial Intelligence Based Algorithms for Prostate Cancer Classification and Detection on Magnetic Resonance Imaging : A Narrative Review ». en. In : *Diagnostics* 11.6 (juin 2021). Number : 6 Publisher : Multidisciplinary Digital Publishing Institute, p. 959. DOI : 10.3390/diagnostics11060959 (cf. p. 42, 43).

[127] Gabriele Valvano, Andrea Leo et Sotirios A. Tsaftaris. « Learning to Segment From Scribbles Using Multi-Scale Adversarial Attention Gates ». In : *IEEE Trans. Med. Imaging* 40.8 (août 2021), p. 1990-2001. ISSN : 0278-0062, 1558-254X. DOI : 10.1109/TMI.2021.3069634 (cf. p. xx, 115).

[128] Sadhna Verma, Baris Turkbey, Naira Muradyan, Arumugam Rajesh, Francois Cornud, Masoom A. Haider, Peter L. Choyke et Mukesh Harisinghani. « Overview of Dynamic Contrast-Enhanced MRI in Prostate Cancer Diagnosis and Management ». In : *American Journal of Roentgenology* 198.6 (juin 2012). Publisher : American Roentgen Ray Society, p. 1277-1288. ISSN : 0361-803X. DOI : 10.2214/AJR.12.8510 (cf. p. 131).

[129] Bo Wang, Yang Lei, Sibo Tian, Tonghe Wang, Yingzi Liu, Pretesh Patel, Ashesh B. Jani, Hui Mao, Walter J. Curran, Tian Liu et Xiaofeng Yang. « Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation ». In : *Medical Physics* 46.4 (avr. 2019), p. 1707-1718. ISSN : 0094-2405. DOI : 10.1002/mp.13416 (cf. p. 50).

[130] Z. Wang, C. Liu, D. Cheng, L. Wang, X. Yang et K. Cheng. « Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network ». In : *IEEE Transactions on Medical Imaging* 37.5 (mai 2018), p. 1127-1139. ISSN : 0278-0062. DOI : 10.1109/TMI.2017.2789181 (cf. p. xx, 44, 45, 51, 116).

[131] Rogier R. Wildeboer, Ruud J. G. van Sloun, Hessel Wijkstra et Massimo Mischi. « Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods ». In : *Computer Methods and Programs in Biomedicine* 189 (2020), p. 105316 (cf. p. 19, 42, 43).

[132] David J. Winkel, Christian Wetterauer, Marc Oliver Matthias, Bin Lou, Bibo Shi, Ali Kamen, Dorin Comaniciu, Hans-Helge Seifert, Cyrill A. Rentsch et Daniel T. Boll. « Autonomous Detection and Classification of PI-RADS Lesions in an MRI Screening Population Incorporating Multicenter-Labeled Deep Learning and Biparametric Imaging : Proof of Concept ». en. In : *Diagnostics* 10.11 (nov. 2020). Number : 11 Publisher : Multidisciplinary Digital Publishing Institute, p. 951. DOI : 10.3390/diagnostics10110951 (cf. p. 53, 146).

[133] Tineke Wolters, Monique J. Roobol, Leeuwen Pim J. van, den Bergh Roderick C. N. van den, Robert F. Hoedemaeker, Leenders Geert J. L. H. van, öder Fritz H. Schr et der Kwast Theodorus H. van der. « A Critical Analysis of the Tumor Volume Threshold for Clinically Insignificant Prostate Cancer Using a Data Set of a Randomized Screening Trial ». In : *Journal of Urology* 185.1 (jan. 2011). Publisher : WoltersKluwer, p. 121-125. DOI : 10.1016/j.juro.2010.08.082 (cf. p. 116).

[134] Sungmin Woo, Chong Hyun Suh, Sang Youn Kim, Jeong Yeon Cho, Seung Hyup Kim et Min Hoan Moon. « Head-to-Head Comparison Between Biparametric and Multiparametric MRI for the Diagnosis of Prostate Cancer : A Systematic Review and Meta-Analysis ». In : *American Journal of Roentgenology* 211.5 (nov. 2018). Publisher : American Roentgen Ray Society, W226-W241. ISSN : 0361-803X. DOI : 10.2214/AJR.18.19880 (cf. p. 131).

[135] Helen Xu, John S. H. Baxter, Oguz Akin et Diego Cantor-Rivera. « Prostate cancer detection using residual networks ». en. In : *Int J CARS* (avr. 2019). ISSN : 1861-6429. DOI : 10.1007/s11548-019-01967-5 (cf. p. 44, 51).

[136] Haibo Yang, GuangYu Wu, Dinggang Shen et Shu Liao. « Automatic Prostate Cancer Detection On Multi-Parametric Mri With Hierarchical Weakly Supervised Learning ». In : *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. ISSN : 1945-8452. Avr. 2021, p. 316-319. DOI : 10.1109/ISBI48211.2021.9434108 (cf. p. 116).

[137] Inwan Yoo, Donggeun Yoo et Kyunghyun Paeng. « PseudoEdgeNet : Nuclei Segmentation only with Point Annotations ». In : *arXiv :1906.02924 [cs]*. arXiv : 1906.02924. Juill. 2019 (cf. p. xix, 114).

[138] Ji Yoon, Moon Choi, Young-Joon Lee et Seung Jung. « Dynamic Contrast-Enhanced MRI of the Prostate : Can Auto-Generated Wash-in Color Map Be Useful in Detecting Focal Lesion Enhancement ? » In : *Investigative Magnetic Resonance Imaging* 23 (jan. 2019), p. 220. DOI : 10.13104/imri.2019.23.3.220 (cf. p. 134).

[139] Fatemeh Zabihollahy, Nicola Schieda, Satheesh Krishna Jeyaraj et Eranga Ukwatta. *Automated segmentation of prostate zonal anatomy on T2-weighted (T2W) and apparent diffusion coefficient (ADC) map MR images using U-Nets*. en. Juill. 2019. DOI : 10.1002/mp.13550 (cf. p. 50, 70).

[140] Olmo Zavala-Romero, Adrian L. Breto, Isaac R. Xu, Yu-Cherng C. Chang, Nicole Gautney, Alan Dal Pra, Matthew C. Abramowitz, Alan Pollack et Radka Stoyanova. « Segmentation of prostate and prostate zones using deep learning ». en. In : *Strahlenther Onkol* 196.10 (oct. 2020), p. 932-942. ISSN : 1439-099X. DOI : 10.1007/s00066-020-01607-x (cf. p. 43, 50, 70, 102).

[141] Guokai Zhang, Weigang Wang, Dinghao Yang, Jihao Luo, Pengcheng He, Yongtong Wang, Ye Luo, Binghui Zhao et Jianwei Lu. « A Bi-Attention Adversarial Network for Prostate Cancer Segmentation ». In : *IEEE Access* 7 (2019), p. 131448-131458. ISSN : 2169-3536. DOI : 10.1109/ACCESS.2019.2939389 (cf. p. 81).

[142] Pengyi Zhang, Yunxin Zhong et Xiaoqiong Li. « ACCL : Adversarial constrained-CNN loss for weakly supervised medical image segmentation ». In : *arXiv :2005.00328 [cs]* (mai 2020). arXiv : 2005.00328 (cf. p. 115).

[143] Jing Zhao, Avan Kader, Dilyana B. Mangarova, Julia Brangsch, Winfried Brenner, Bernd Hamm et Marcus R. Makowski. « Dynamic Contrast-Enhanced MRI of Prostate Lesions of Simultaneous [68Ga]Ga-PSMA-11 PET/MRI : Comparison between Intraprostatic Lesions and Correlation between Perfusion Parameters ». en. In : *Cancers* 13.6 (jan. 2021). Number : 6 Publisher : Multidisciplinary Digital Publishing Institute, p. 1404. DOI : 10.3390/cancers13061404 (cf. p. 143).

[144] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva et A. Torralba. « Learning Deep Features for Discriminative Localization ». In : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN : 1063-6919. Juin 2016, p. 2921-2929. DOI : 10.1109/CVPR.2016.319 (cf. p. xix, 109, 110).

[145] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh et J. Liang. « UNet++ : Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation ». In : *IEEE Transactions on Medical Imaging* 39.6 (juin 2020). Conference Name : IEEE Transactions on Medical Imaging, p. 1856-1867. ISSN : 1558-254X. DOI : 10.1109/TMI.2019.2959609 (cf. p. 81).

[146] Qikui Zhu, Bo Du et Pingkun Yan. « Boundary-Weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation ». In : *IEEE Transactions on Medical Imaging* 39.3 (mars 2020). Conference Name : IEEE Transactions on Medical Imaging, p. 753-763. ISSN : 1558-254X. DOI : 10.1109/TMI.2019.2935018 (cf. p. 50).

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : DURAN
(avec précision du nom de jeune fille, le cas échéant)

DATE de SOUTENANCE : 03/02/2022

Prénoms : Audrey Marie-Anne

TITRE : Intelligence artificielle pour la caractérisation du cancer de la prostate par agressivité en IRM multiparamétrique

NATURE : Doctorat

Numéro d'ordre : 2022LYSEI008

Ecole doctorale : Électronique, Électrotechnique, Automatique (EEA)

Spécialité :Traitement du Signal et de l'Image

RESUME : Le cancer de la prostate (CaP) est le cancer le plus diagnostiqué dans plus de la moitié des pays du monde et le cinquième cancer le plus meurtrier chez les hommes en 2020. Le diagnostic du CaP inclut une acquisition IRM multiparamétrique (IRM-mp, qui combine les séquences T2-w, DWI et DCE) avant même la réalisation de biopsies. Toutefois, l'analyse jointe de ces images multimodales est fastidieuse et chronophage, en particulier lorsque les séquences mènent à des conclusions différentes. En outre, la sensibilité de l'IRM reste faible pour les cancers peu agressifs et la variabilité inter-observateur élevée. De plus, l'analyse visuelle ne permet pas aujourd'hui de déterminer l'agressivité des cancers, caractérisée par le score de Gleason (GS).
C'est pourquoi de nombreux systèmes binaires d'aide au diagnostic (CAD) basés sur des modèles statistiques par apprentissage ont été proposés ces dernières années, dans le but d'assister le radiologue à détecter les lésions cliniquement significatives (CS). L'objectif de cette thèse est d'élaborer un système CAD pour détecter les CaP sur des images IRM-mp mais aussi caractériser leur agressivité en prédisant le GS associé.

Dans une première partie, nous présentons un système CAD supervisé permettant de détecter et caractériser le CaP par agressivité à partir des cartes T2-w et ADC. Ce réseau de neurones a la particularité d'utiliser la prostate comme carte d'attention pour mieux caractériser les lésions présentes et il obtient les meilleures performances de l'état de l'art pour la caractérisation du score de Gleason. Ce modèle basé sur un réseau de neurones nécessite une base de données importante, difficile à obtenir et fastidieuse à annoter. C'est pourquoi, dans une deuxième partie, nous nous penchons sur un modèle faiblement supervisé, permettant l'inclusion de données où les lésions sont identifiées par des points seulement, pour un gain de temps conséquent et l'inclusion de bases de données établies sur la biopsie. Dans une dernière partie, nous étudions l'apport de l'imagerie DCE dans un système CAD pour la caractérisation du CaP, séquence souvent omise en entrée des modèles profonds.

MOTS-CLÉS : segmentation sémantique, apprentissage profond, cancer de la prostate, systèmes d'aide au diagnostic, imagerie par résonance magnétique, modèles d'attention, apprentissage faiblement supervisé

Laboratoire (s) de recherche : Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé (CREATIS)

Directeur de thèse: Carole LARTIZIEN

Président de jury : Caroline PETITJEAN

Composition du jury :

Mériaudeau, Fabrice	Professeur des Universités	Université de Bourgogne	Rapporteur
Petitjean, Caroline	Professeure des Universités	Université de Rouen	Rapporteure
Jodoin, Pierre-Marc	Professeur des Universités	Université de Sherbrooke	Examinateur
Renard-Penna, Raphaële	Professeure des Universités-Praticien Hospitalier	Sorbonne Universités	Examinatrice
Rouvière, Olivier	Professeur des Universités-Praticien Hospitalier	Université Lyon 1	Examinateur
Lartizien, Carole	Directrice de recherche	CNRS, Lyon	Directrice de thèse