# Segmentation et identification des vertèbres sur des radiographies frontales de rachis par méthodes d'apprentissage profond

par

Agathe MONCORGÉ

RAPPORT DE MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAITRISE EN GÉNIE
CONCENTRATION PERSONNALISÉE
M. Sc. A.

MONTRÉAL, LE 23 FÉVRIER 2022

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

 Agathe Moncorgé, 2022

Cette licence Creative Commons signifie qu'il est permis de diffuser, d'imprimer ou de sauvegarder sur un autre support une partie ou la totalité de cette œuvre à condition de mentionner l'auteur, que ces utilisations soient faites à des fins non commerciales et que le contenu de l'œuvre n'ait pas été modifié.

**PRÉSENTATION DU JURY**

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Carlos Vázquez, directeur de mémoire
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Jacques de Guise, codirecteur de mémoire
Département de génie des systèmes à l'École de technologie supérieure

M. Matthew Toews, président du jury
Département de génie des systèmes à l'École de technologie supérieure

M. José Dolz, membre du jury
Département de génie logiciel à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 17 FÉVRIER 2022

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

## REMERCIEMENTS

J’aimerais tout d’abord remercier mon directeur de recherche, Carlos, pour sa confiance, sa bienveillance et ses précieux conseils qui ont guidé la réalisation de mon projet.

Je remercie aussi mon co-directeur Jacques pour son chaleureux accueil au LIO, sa bienveillance et son éthique de travail. Sans le cours GTS815, qui a révélé mon intérêt pour l’imagerie médicale, ce projet n’aurait jamais eu lieu. Bien sûr, je remercie Thierry pour son soutien, ses réponses à mes innombrables questions et son implication auprès des étudiants, notamment pendant la pandémie.

Je remercie aussi l’entreprise EOS Imaging et le CRSNG, sans qui ce projet n’aurait pas été possible.

Merci beaucoup à tous mes collègues du LIO, notamment Marie, Camille, Manon, Hélène, Victor, Francis, Stevy... J’ai beaucoup apprécié le temps que j’ai passé avec vous au labo (et sur Zoom), même s’il a été quelque peu écourté. J’espère bien vous revoir par la suite !

Merci à tous les colocs de cette aventure québécoise : Ali, Tutur, Sims, Keks, Anthony, Axel, Valentin, Sonia, Mathéo et Megan. Ça aurait été vachement moins le fun sans vous. Évidemment, je remercie les Phy’stons et ma gang de ministres. C’est quand que vous (re-) venez me voir au Québec ? Je vous attends.

Mille mercis à mes parents pour leur soutien sans faille et leurs encouragements ; à mes grands-parents pour leurs appels réconfortants; à Gen pour la correspondance épistolaire longue de deux belles années et à mes petits frères, qui savent toujours me distraire. Enfin, je remercie mon bien-aimé Alexis, pour tout.

**Segmentation et identification des vertèbres sur des radiographies frontales de rachis par méthodes d'apprentissage profond**

Agathe MONCORGÉ

## RÉSUMÉ

Les outils d'extraction de paramètres cliniques et de reconstruction 3D du rachis à partir de radiographies aident les cliniciens dans le diagnostic et le traitement des patients atteints de scolioses. L'entreprise EOS Imaging cherche à automatiser ces applications afin de mieux les intégrer en routine clinique. La segmentation et l'identification des vertèbres sur les radiographies sont deux étapes fondamentales et cruciales dans les processus d'extraction de paramètres cliniques et de reconstruction 3D du rachis. De ce fait, l'automatisation de ces étapes relativement complexes nécessite une attention particulière.

L'objectif principal de ce projet est de proposer une méthode automatique de segmentation et de labellisation individuelle des vertèbres à partir des radiographies EOS frontales de rachis atteints de scoliose idiopathique de l'adolescent (SIA). Le projet se tourne vers les méthodes d'apprentissage profond, spécifiquement les réseaux de neurones convolutifs (CNN).

Une base de données de 767 images radiographiques frontales de patients atteints de SIA est utilisée. Un soin particulier est apporté au choix des métriques d'évaluation afin que ces dernières reflètent bien les capacités des CNNs à produire des résultats utilisables dans un contexte clinique. Une étude comparative des CNNs de segmentation d'instance DetectoRS, RetinaMask, MS R-CNN et YOLACT est réalisée. DetectoRS est significativement meilleur que les trois autres CNNs sur la métrique de taux d'identification la plus contraignante.

La méthode proposée s'appelle V-DetectoRS. Le CNN V-DetectoRS est construit à partir de DetectoRS et intègre deux stratégies d'ajustement basées sur le respect de la structure anatomique du rachis. Un post-traitement ainsi qu'un terme de pénalité intégré à la fonction de perte en régression de DetectoRS sont mis en place. Ils assurent tous les deux le respect des distances intervertébrales. Les performances atteintes par V-DetectoRS surpassent celles de DetectoRS sur toutes les métriques d'évaluation. Les performances obtenues avec V-DetectoRS sont similaires à celles des méthodes de la revue de littérature. V-DetectoRS atteint

un coefficient de Dice moyen de 87,84%. En considérant qu'une vertèbre est bien identifiée lorsqu'elle montre un coefficient de Dice de 75%, 86,21% des images testées ont été traitées avec succès par V-DetectoRS.
Dans le futur, V-DetectoRS pourrait être intégré à des applications de reconstruction 3D de rachis ou d'extraction de paramètres cliniques afin de participer à leur automatisation.

**Mots-clés :** réseau neuronal convolutif, segmentation d'instance, vertèbres, radiographie

**Segmentation and identification of vertebrae on frontal X-rays of the spine with deep learning methods**

Agathe MONCORGÉ

## ABSTRACT

Clinical parameters extraction and 3D spine reconstruction tools from X-rays help clinicians in the diagnosis and treatment of scoliotic patients. The company EOS Imaging aims to automate such applications to better include them in clinical routine. Vertebrae segmentation and identification on X-rays represent a fundamental and crucial step in the clinical parameters and 3D spine reconstruction processes. Hence, the automatization of this relatively complex step requires particular attention.

The main goal of this project is to propose an automatic method to segment and individually identify vertebrae from EOS frontal X-rays of the spine. The project leverage deep learning methods, convolutional neural networks more specifically.

A database of 767 frontal X-rays from patients suffering from idiopathic adolescent scoliosis is used. Careful attention was paid to the choice of evaluation metrics so that they reflect the CNNs ability to produce clinically usable results. A comparative study of the instance segmentation CNNs DetectoRS, RetinaMask, MS R-CNN et YOLACT is conducted. DetectoRS is significantly better than all three others CNNs on the most constraining identification rate metric.

The proposed method is called V-DetectoRS. The V-DetectoRS CNN is built on DetectoRS and incorporates two adjustment strategies based on the respect of the anatomic structure of the spine. A post-treatment and a penalty term added to the regression loss function of DetectoRS are added. They both ensure the respect of intervertebral distances. V-DetectoRS's results outperform DetectoRS on every evaluation metric. Compared with the literature review methods, V-Detectors shows similar performances. V-DetectoRS reaches a mean Dice coefficient of 87.84%. Considering that a vertebra is well identified when the Dice coefficient is above 75%, 86.21% of the tested images are processed with success by V-DetectoRS.

In the future, V-Detectors could be incorporated in clinical parameters and 3D spine reconstruction applications in order to participate in their automatization.

**Keywords:** convolutional neural network, instance segmentation, vertebrae, radiography

## TABLE DES MATIÈRES

Page

## LISTE DES TABLEAUX

## LISTE DES FIGURES

## LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

| | |
|---|---|
| 2D | Bidimensionnel |
| 3D | Tridimensionnel |
| AP | Average Precision – Précision moyenne |
| ASPP | Atrous Spatial Pyramid Pooling – Mise en commun pyramidale spatiale à trous |
| C1 à C7 | Première vertèbre cervicale à la septième |
| CE | Cross Entropy – Entropie croisée |
| CNN | Convolutional Neural Network – Réseau de neurones convolutif |
| CT-scan | Computed Tomography scan – Scan tomodensitométrique |
| DR-Nets | Deep Reasoning Networks – Réseaux de raisonnement profond |
| DSC | Dice Similarity Coefficient – Coefficient de similarité de Dice |
| EOS | Système d'imagerie médical, EOS Imaging, France |
| FC | Fully Connected – Entièrement connecté |
| FCN | Fully Convolutional Network – Réseau entièrement convolutif |
| FN | Faux négatif |
| FP | Faux positif |
| VP | Vrai positif |
| FPN | Feature Pyramid Networks – Réseau pyramidal de caractéristiques |
| HTC | Hybrid Task Cascade – Tâche hybride en cascade |
| IoU | Intersection over Union – Indice de Jaccard |
| IR | Identification Rate – Taux d'identification |
| L1 à L5 | Première vertèbre lombaire à la cinquième |

| | |
|---|---|
| MAE | Mean Absolute Error – Moyenne des erreurs absolues |
| MICCAI | Medical Image Computing and Computer Assisted Intervention |
| MIoU | Mean Intersection over Union – Indice de Jaccard moyen |
| MLP | Multi Layers Perceptron – Perceptron à multiples couches |
| MS-RCNN | Mask Scoring R-CNN |
| MURA | Musculoskeletal radiographs – Radiographies musculo-squelettiques |
| NMS | Non-Maximum Suppression – Suppression non maximum |
| PSPNet | Pyramid Scene Parsing Network – Réseau d'analyse de scène pyramidale |
| R-CNN | Region based Convolutional Neural Network – Réseau de neurones convolutif basé sur les régions |
| ReLU | Rectified Linear Unit – Unité de rectification linéaire |
| ResNet | Residual Network – Réseau résiduel |
| RFP | Recursive Pyramid Feature – Caractéristiques de pyramide récursive |
| RoIs | Regions of Interest – Régions d'intérêt |
| RPN | Regions Proposal Network – Réseau de proposition de régions |
| S1 à S5 | Première vertèbre sacrée à la cinquième |
| SAC | Switchable Atrous Convolution – Convolution commutable à trous |
| SGD | Stochastic Gradient Descent – Descente de gradient stochastique |
| SIA | Scoliose idiopathique de l'adolescent |
| SCN | SpatialConfiguration-Net |
| SOSORT | Scientific Society on Scoliosis Orthopaedic and Rehabilitation Treatment – Société scientifique sur le traitement orthopédique et de réadaptation de la scoliose |

| | |
|---|---|
| SR | Success Rate – Taux de succès |
| SterEOS© | Logiciel de reconstruction 3D commercial (EOS Imaging, Paris, France) |
| SVM | Support Vector Machine – Machine à vecteurs de support |
| T1 à T12 | Première vertèbre thoracique à la douzième |
| VerSe | Large Scale Vertebrae Segmentation Challenge – Défi de segmentation des vertèbres à grande échelle |
| YOLACT | You Only Look At CoefficienTs |

## LISTE DES SYMBOLES ET UNITÉS DE MESURE

UNITÉS GÉOMÉTRIQUES

**Angle plan**

° degré

**Longueur**

mm millimètre

# INTRODUCTION

La scoliose idiopathique de l'adolescent, qui consiste en une déformation tridimensionnelle de la colonne vertébrale, touche entre 3 et 4% de la population mondiale (Konieczny, Senyurt et Krauspe, 2013; Choudhry, Ahmad et Verma, 2016). Il existe aujourd'hui différents outils qui aident les cliniciens dans le diagnostic et le traitement de la scoliose. L'entreprise EOS Imaging propose notamment des outils pour extraire des métriques cliniques et reconstruire des modèles 3D du rachis depuis des images radiographiques biplanes. Pour s'affranchir du besoin d'opérateur et ainsi mieux intégrer ces outils en routine clinique, EOS Imaging cherche à automatiser ces derniers.

Ce projet s'inscrit dans le cadre de l'automatisation des applications de l'entreprise EOS Imaging, plus spécifiquement dans la segmentation et l'identification des vertèbres sur des radiographies.

Afin de saisir les enjeux de ce projet, l'introduction propose de brefs rappels sur l'anatomie du rachis, sur la scoliose et son diagnostic et sur les intérêts et défi liés à l'automatisation des outils développés par EOS Imaging. L'introduction présente aussi l'objectif du projet ainsi que la structure du mémoire.

## 0.1 Rachis et scoliose

Ce projet traitant de la segmentation et de l'identification de vertèbres sur des radiographies, il est nécessaire de rappeler des notions de base quant à l'anatomie du rachis et à la scoliose.

### 0.1.1 Anatomie du rachis

La colonne vertébrale, aussi appelée rachis, est une structure appartenant au squelette du tronc. Le rachis comprend généralement 33 vertèbres : 7 vertèbres cervicales (notées C1 à C7), 12 vertèbres thoraciques (notées T1 à T12), 5 vertèbres lombaires (notées L1 à L5), 5 vertèbres sacrées qui sont fusionnées et qui forment le sacrum (S1 à S5) et 4 vertèbres coccygiennes qui

sont soudées et qui forment le coccyx (Moulin, 2014 ; Lamarre, 2008). Une représentation des éléments osseux du rachis est donnée en Figure 0.1.

Dans le plan frontal, les vertèbres d'un rachis sain suivent un arrangement vertical. Dans le plan sagittal, on observe quatre courbures naturelles : la lordose cervicale, la cyphose thoracique, la lordose lombaire et la cyphose sacrée (Moulin, 2014 ; Lamarre, 2008).

La partie antérieure d'une vertèbre, en forme de cylindre, est le corps vertébral. Il présente à chacune de ses extrémités supérieures et inférieures un plateau vertébral. En arrière du corps vertébral se situe un arc neural composé de deux pédicules, deux lames et sept apophyses (2 transverses, 1 épineuse et 4 articulaires). Le schéma anatomique d'une vertèbre est donné en Figure 0.2.

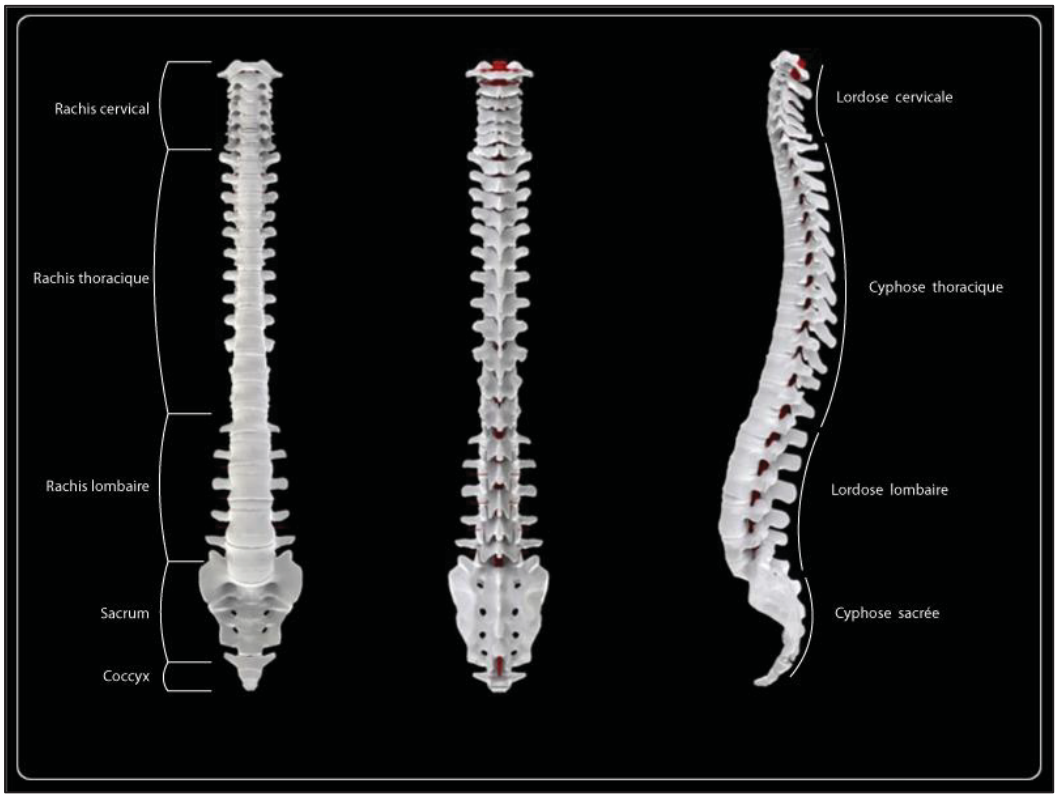



Figure 0.1 La colonne vertébrale en vue frontale et latérale – Anatomie
Tirée de Kaudris, CC BY-SA 3.0, via Wikimedia Commons

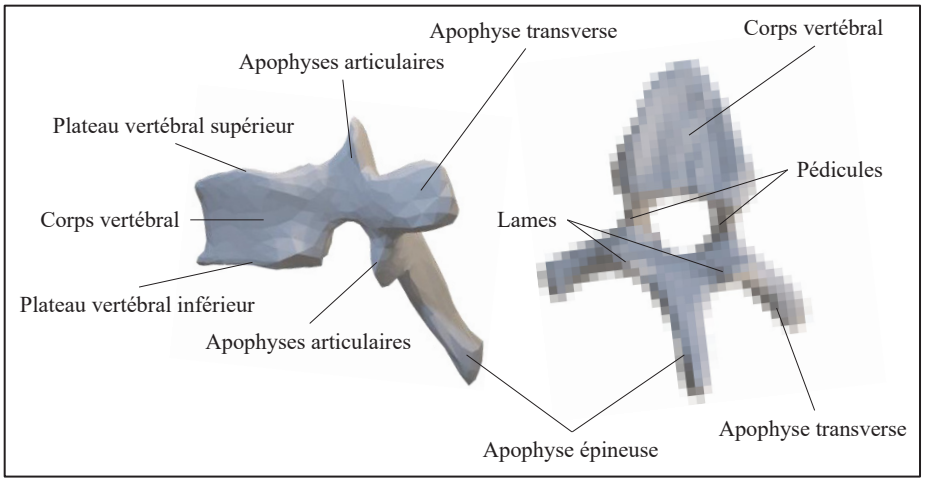



Figure 0.2 Schéma anatomique d'une vertèbre

### 0.1.2 La scoliose

La scoliose consiste en une déformation tridimensionnelle de la colonne vertébrale. Cette déformation provient d'une torsion généralisée du rachis due à des déplacements de vertèbres (Lamarre, 2008). Une radiographie frontale d'un rachis scoliotique est montrée à la Figure 0.3. 80% des cas de scolioses sont définis comme des scolioses idiopathiques de l'adolescent (SIA), c'est-à-dire des scolioses qui touchent les enfants et adolescents (principalement les filles) pendant leur croissance et dont les principales causes sont inconnues (Negrini *et al.*, 2012 ; Choudhry, Ahmad et Verma, 2016). Les scolioses idiopathiques de l'adolescent sont fréquentes, car on estime qu'entre 3 et 4% de la population mondiale est affectée (Konieczny, Senyurt et Krauspe, 2013; Choudhry, Ahmad et Verma, 2016). De nombreux travaux ont posés différentes hypothèses quant à l'étiologie de la scoliose idiopathique : des prédispositions héréditaires ou génétiques, des malformations congénitales, un déficit en mélatonine (Machida, 1999 ; Weinstein *et al.*, 2008) ou une asymétrie vestibulaire pourraient expliquer l'apparition de scolioses chez l'adolescent (Lambert *et al.*, 2009).

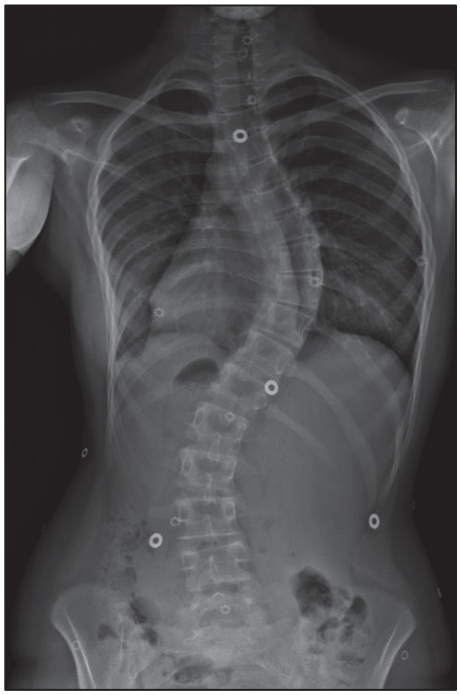


Figure 0.3 Radiographie frontale
de rachis scoliotique

Afin de diagnostiquer la scoliose, les courbures du rachis sont caractérisées à l'aide de différentes observations et mesures. Une courbure se décrit en identifiant trois vertèbres caractéristiques. La première est la vertèbre apicale, vertèbre de la courbure la plus latéralement déviée et montrant une rotation vertébrale axiale importante. Les deux autres sont les vertèbres dites limites inférieures et supérieures, qui sont les points d'inflexion des courbures. On les reconnait à leur inclinaison par rapport à l'horizontale plus prononcée que pour les autres vertèbres (Moulin, 2014). L'angle de Cobb est la métrique clinique de référence, utilisée par les cliniciens pour quantifier l'amplitude des courbures scoliotiques d'un rachis (Moulin, 2014 ; Lamarre, 2008). Comme montré sur la Figure 0.4, il correspond à l'angle formé par l'intersection de la tangente au plateau supérieur de la vertèbre limite supérieure et de la tangente au plateau inférieur de la vertèbre limite inférieure. L'angle de Cobb est généralement directement mesuré sur la radiographie frontale du patient (Labelle *et al.*, 2011). Il est reconnu que l'erreur de mesure maximale acceptable de l'angle de Cobb est de 5° (Negrini *et al.*, 2012).

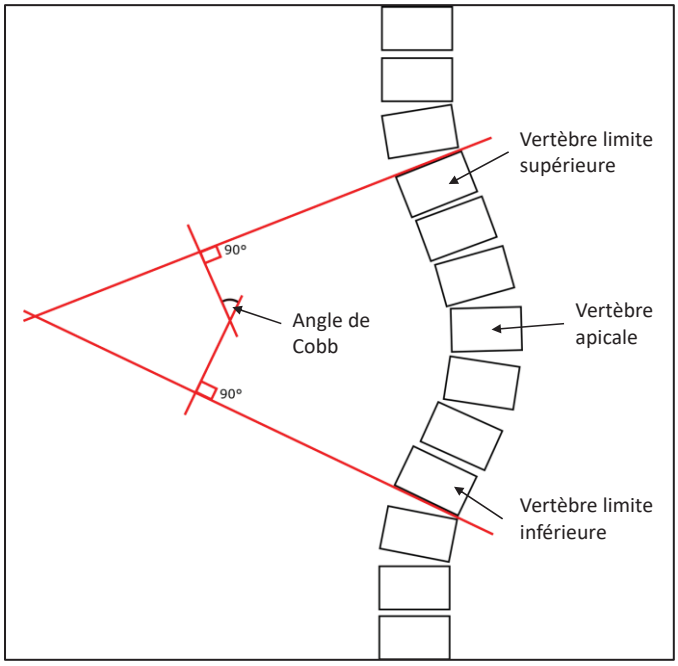



Figure 0.4 Schéma de calcul de l'angle de Cobb

Les paramètres cliniques sont généralement évalués directement sur les radiographies frontales et latérales (Illés, Lavaste et Dubousset, 2019). Cependant, ces mesures 2D ne reflètent pas le caractère tridimensionnel de la scoliose. Le Dr Labelle (Labelle *et al.*, 2011) souligne ainsi l'importance de la reconstruction 3D du rachis. La visualisation de la scoliose en trois dimensions améliore l'appréhension des déformations 3D, la définition des objectifs de traitement, l'analyse des résultats de traitements et la reproductibilité de classification des scolioses (Labelle *et al.*, 2011).

## 0.2 EOS Imaging

Ce projet de recherche est réalisé en collaboration avec les partenaires industriels EOS Imaging (Paris, France) et sa filiale Montréalaise EOS Image Inc. (Canada), qui développent des solutions d'imagerie et d'orthopédie.

### 0.2.1 Solutions d'imagerie et d'analyse d'images

Le système EOS® (EOS Imaging, Paris, France) est un équipement stéréoradiographique à basse dose permettant l'acquisition simultanée de deux radiographies biplanaires du corps entier. Le système, basé sur les travaux de Georges Charpak (Prix Nobel de Physique en 1992), réduit considérablement la dose de radiation par rapport aux systèmes radiographiques standard. De plus, l'acquisition simultanée des radiographies calibrées dans les plans frontal et latéral rend possible l'analyse 3D de la scoliose (EOS imaging, 2021).
EOS Imaging développe aussi des outils d'analyse des radiographies, notamment pour assister le clinicien dans le diagnostic et le choix de traitement de la scoliose. Le logiciel SterEOS© est une plateforme logiciel qui permet entre autres de reconstruire des modèles 3D personnalisés du rachis et d'estimer des paramètres cliniques associés 2D et 3D tels que l'angle de Cobb ou la rotation axiale vertébrale à partir des radiographies biplanaires acquises avec le système EOS®.

### 0.2.2 Automatisation des outils

Un des objectifs d'EOS Imaging est l'automatisation des outils d'analyse des radiographies. À ce jour, l'outil dédié à la reconstruction 3D du rachis disponible dans l'application commerciale SterEOS©, qui repose sur les travaux de Humbert et al. (Humbert *et al.*, 2009), est semi-automatique. Cette méthode de reconstruction nécessite des ajustements manuels complexes et peu intuitifs d'un modèle 3D de la colonne vertébrale : suivant le niveau d'expertise de l'opérateur et du type de scoliose, la reconstruction peut prendre jusqu'à 40 minutes et montrer une large variabilité interopérateurs. Ces limites expliquent la faible intégration de la méthode dans les routines cliniques (Aubert, 2020).
L'automatisation complète des outils, permettant de s'affranchir du besoin d'opérateur et d'accélérer les processus, montre ainsi un réel intérêt. (Aubert, 2020) propose dans sa thèse une méthode de reconstruction 3D du rachis complètement automatique à partir des radiographies EOS biplanaires. Elle n'existe cependant pas encore sous version commerciale.

Les outils développés par EOS Imaging -que ce soit l'outil de reconstruction 3D de rachis ou de calcul de paramètres cliniques- passent par une phase fondamentale d'extraction d'informations anatomiques depuis les images radiographiques (Humbert *et al.*, 2009 ; Aubert, 2020). Cette étape cruciale, qui doit être réalisée avec justesse et fiabilité, est relativement difficile à automatiser (Aubert, 2020).

Les informations anatomiques à extraire depuis les radiographies requises pour l'initialisation et l'affinement du modèle 3D du rachis consistent en l'identification et la segmentation de chacune des vertèbres (Aubert, 2020). L'angle de Cobb, paramètre clinique d'évaluation de la scoliose de référence, est déterminé avec la radiographie en vue frontale seulement. Plus spécifiquement, les informations d'identification des vertèbres, mais aussi la localisation et l'orientation des plateaux vertébraux sont nécessaires pour calculer l'angle de Cobb. Des exemples de masques de segmentation des vertèbres entières (soit le corps vertébral, les lames et les apophyses) sont donnés en Figure 0.5.

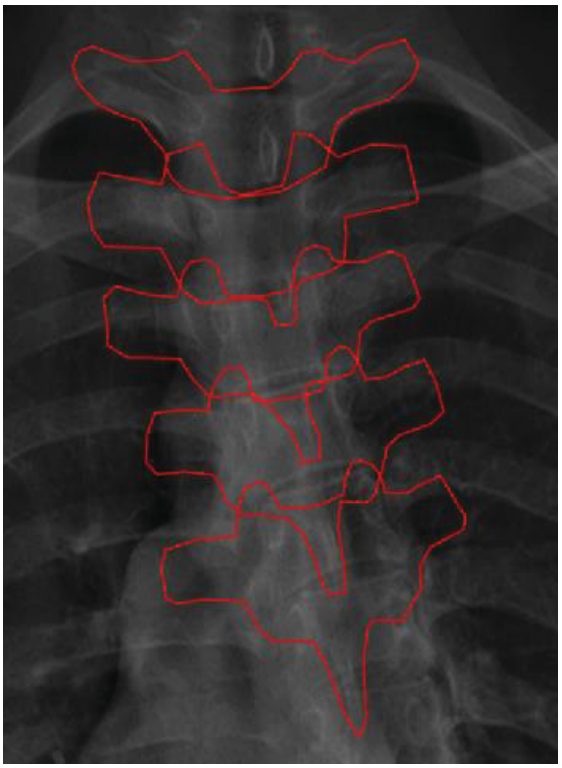


Figure 0.5 Exemples de masques de segmentation de vertèbres sur une radiographie EOS en vue frontale

## 0.3 Objectif général du projet

L'automatisation de l'extraction de repères anatomiques depuis les radiographies constitue une étape complexe et fondamentale dans le processus d'automatisation des applications EOS d'analyse de la scoliose.

Ce projet est défini à partir de ce constat. L'objectif général est de proposer une méthode automatique de segmentation et de labellisation individuelle des vertèbres depuis des radiographies EOS frontales de rachis atteints de SIA. Ce projet constitue un travail exploratoire qui participe au processus d'automatisation des outils de reconstruction 3D du rachis et d'évaluation des paramètres cliniques d'EOS Imaging.

## 0.4 Structure du document

Cette introduction est suivie par la revue de littérature, en CHAPITRE 1. À l'issue de la revue, la problématique ainsi que les objectifs du projet sont fixés au CHAPITRE 2. Une étude comparative d'algorithmes d'apprentissage profond pour la tâche d'identification et de segmentation des vertèbres sur des radiographies est présentée au CHAPITRE 3. Le CHAPITRE 4 constitue une proposition d'amélioration d'un algorithme d'apprentissage profond. Le document se termine par une discussion générale et une conclusion sur le travail réalisé.

# CHAPITRE 1

# REVUE DE LA LITTÉRATURE

Les techniques d'apprentissage automatique, notamment d'apprentissage profond, ont révolutionné le domaine de la vision par ordinateur en élargissant le champ des possibles. Elles sont capables d'effectuer des tâches de prédiction jusque-là irréalisables avec des méthodes de vision par ordinateur classiques. Plus spécifiquement, les réseaux de neurones convolutifs (CNNs pour *Convolutional Neural Networks*) sont capables de réaliser avec justesse des tâches de classification d'images, de détection d'objet et de segmentation (Walsh *et al.*, 2019).

L'engouement scientifique pour les CNNs depuis ces dernières années résulte en une amélioration constante des performances des CNNs et à une utilisation étendue à de vastes domaines (Dhillon et Verma, 2019).

La revue de la littérature se tourne donc naturellement vers les méthodes d'apprentissage profond, plus particulièrement les CNNs, pour résoudre le problème d'extraction de repères anatomiques depuis des images radiographiques.

Tout d'abord, des rappels sur les réseaux de neurones, plus particulièrement les réseaux de neurones convolutifs sont faits. Ensuite, une deuxième partie est consacrée à la segmentation et l'identification d'objets sur des images par CNNs. Dans une troisième partie, la revue se spécifie sur la segmentation d'instance de vertèbres sur des images radiologiques. Enfin, différentes stratégies de spécialisation d'un CNN à son cadre d'utilisation sont abordées dans une quatrième et dernière partie.

## 1.1 Notions de base sur les réseaux de neurones convolutifs

Afin de saisir le principe de réseau neuronal convolutif, il est nécessaire de revenir sur le principe l'apprentissage automatique, de réseau de neurones et d'apprentissage profond.

### 1.1.1 L'apprentissage automatique

Les réseaux de neurones font partie des techniques d'apprentissage automatique qui appartiennent elles-mêmes au domaine de l'intelligence artificielle (Hosch, 2021).

L'apprentissage automatique vise à donner la capacité à des systèmes d'apprendre des modèles mathématiques à partir de données et non à l'aide d'instructions explicites (Ongsulee, 2017). Cette compréhension implicite permet de résoudre des tâches trop complexes à programmer explicitement. Chaque système d'apprentissage automatique possède des paramètres et des hyperparamètres. Alors que les paramètres du système sont appris, les valeurs des hyperparamètres doivent être fixées et contrôlent le processus d'apprentissage.

Il existe cinq principaux types d'apprentissages automatiques (Ongsulee, 2017) : l'apprentissage supervisé, l'apprentissage semi-supervisé, l'apprentissage partiellement supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. En apprentissage supervisé, on dispose des informations de sortie, appelées données de vérité terrain, correspondantes à chaque donnée d'entrée. L'objectif est d'apprendre le modèle mathématique qui, pour chaque donnée d'entrée, prédit la sortie associée. Le processus d'apprentissage est qualifié de supervisé, car, pour chaque prédiction émise par le modèle, ce dernier est corrigé en fonction du décalage entre la sortie prédite et la donnée de vérité terrain (Ongsulee, 2017). En apprentissage semi-supervisé, les données de vérité terrain ne sont pas disponibles pour une partie des données tandis qu'elles sont de moindre qualité ou ne sont pas complètes pour l'apprentissage partiellement supervisé (Yan *et al.*, 2017). L'apprentissage non supervisé ne dispose d'aucune donnée de vérité terrain : le modèle cherche à acquérir la structure relative aux données d'entrée. En apprentissage par renforcement, l'apprentissage est guidé par les réponses données par l'environnement à la suite de chaque action entreprise par le modèle (Ongsulee, 2017).

L'apprentissage supervisé est très utilisé en apprentissage machine, il représente environ 70% des algorithmes développés (Ongsulee, 2017). Pour cette raison, la suite de la revue se focalise sur les techniques d'apprentissage supervisé.

### 1.1.2 Les réseaux de neurones

À l'origine, le principe du réseau de neurones artificiels est inspiré du fonctionnement du cerveau humain. En 1957, Frank Rosenblatt (Rosenblatt, 1958) propose le modèle du perceptron en se reposant sur le modèle d'un neurone artificiel de McCulloch-Pitts. Le perceptron, se composant d'un neurone artificiel et d'une fonction d'activation, forme un réseau monocouche capable d'apprendre un simple classificateur linéaire (voir Figure 1.1).

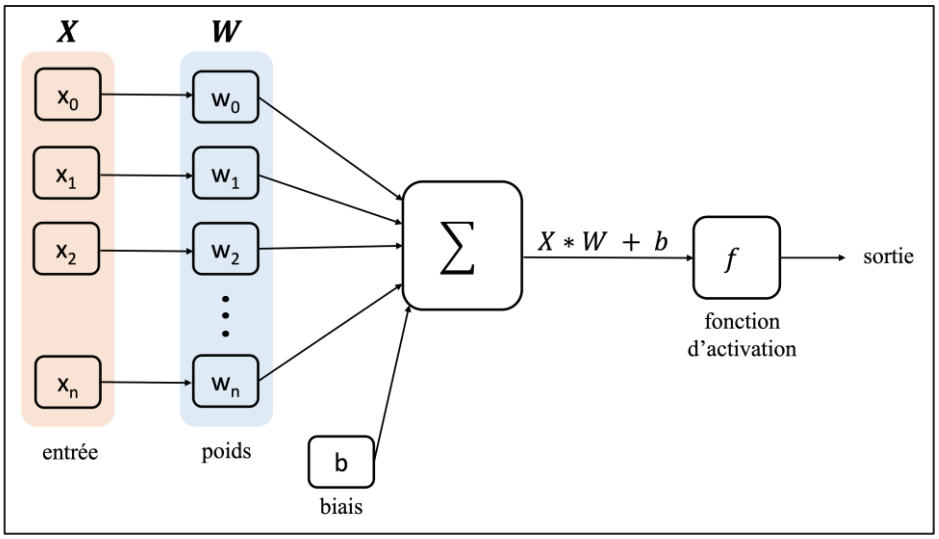



Figure 1.1 Perceptron monocouche

Le neurone prend en entrée un vecteur $X$ de taille n, un vecteur $W$ de poids associés à chacune des entrées ainsi qu'un biais $b$. Une fonction d'activation $f$ génère un signal de sortie à partir de la somme pondérée des entrées calculée par le neurone. La fonction d'activation est non linéaire, c'est généralement une fonction tangente hyperbolique $f(a) = \tanh(a)$, une fonction sigmoïde $f(a) = \frac{1}{1+e^{-a}}$ , ou une fonction de rectification linéaire (ReLU pour *Rectified Linear Unit*) $f(a) = max(0, a)$. Le biais $b$ est une constante qui permet de décaler l'ensemble de la courbe de fonction d'activation afin de correspondre au mieux aux données.
Le perceptron multicouche (MLP pour *Multi Layers Perceptron*) est développé pour résoudre des problèmes non linéaires (Taud et Mas, 2018). C'est le tout premier réseau de neurones à propagation vers l'avant entièrement connecté. Il se compose d'au moins 3 couches : une

couche d'entrée, un certain nombre de couches cachées et une couche de sortie. Les couches sont reliées les unes aux autres par des connexions entre les neurones qui les composent. L'information circule vers l'avant, de la couche d'entrée à la couche de sortie (Taud et Mas, 2018 ; Marius *et al.*, 2009). Une représentation du MLP à deux couches cachées est donnée en Figure 1.2.

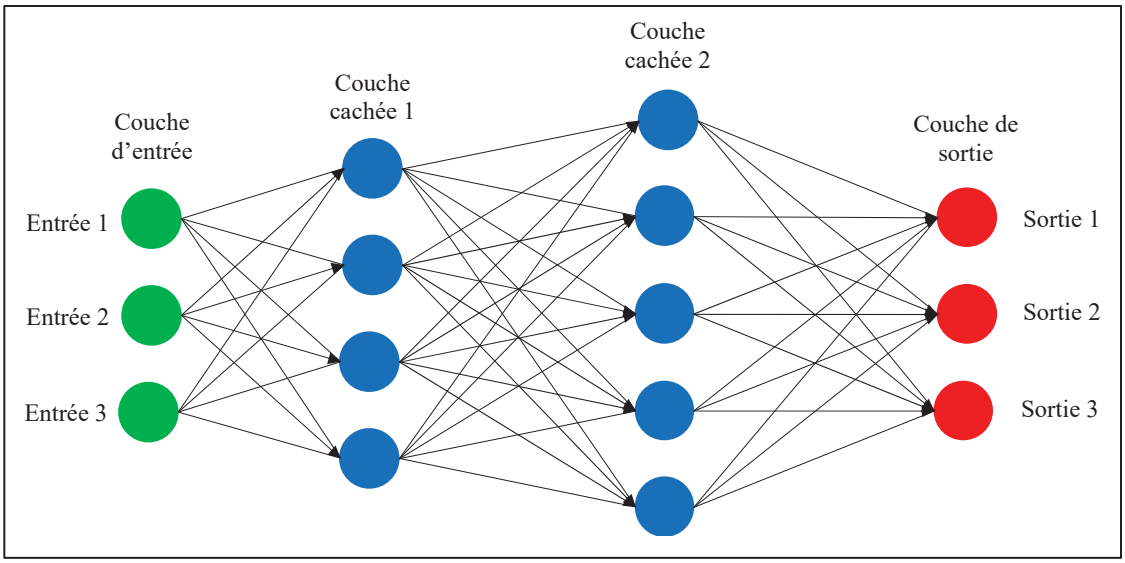



Figure 1.2 Schéma du MLP

Le MLP utilise le concept de rétropropagation du gradient afin de résoudre des problèmes non linéaires en apprentissage supervisé (Marius *et al.*, 2009). Il fonctionne de manière itérative. À chaque itération, un lot d'échantillons, soit de multiples vecteurs d'entrées, est traité par le réseau. Chaque vecteur est propagé vers l'avant à travers les couches du réseau jusqu'à atteindre la couche de sortie, où un vecteur de sortie est obtenu.

Soit $W_i$ le vecteur des poids constituant les couches du MLP à l'itération $i$. La sortie $Y_i$ obtenue est une fonction de $W_i$. À l'issue de cette itération $i$, une fonction de perte $F$ calcule alors une erreur $E_i$ correspondant à la différence entre la sortie obtenue $Y_i(W_i)$ et la sortie attendue $Y'$ .

$$E_i = F(Y_i, Y') \tag{1.1}$$

Le gradient des erreurs $\nabla_{W_i} E_i(W_i)$ est alors calculé par rapport à chaque poids du modèle grâce à l'algorithme de rétropropagation. Pour minimiser la fonction de perte, les paramètres sont actualisés à chaque itération suivant l'algorithme de descente de gradient. L'équation (1.2) présente l'actualisation des poids $W$ du réseau à l'itération $i$.

$$W_{i+1} \leftarrow W_i - \eta \cdot \nabla_{W_i} E_i(W_i) \tag{1.2}$$

À l'itération $i + 1$, le réseau utilise les poids actualisés $W_{i+1}$ pour calculer la sortie $Y_{i+1}$. Le taux d'apprentissage η (*learning rate*) contrôle l'intensité avec laquelle le modèle est modifié à chaque itération. L'important rôle de cet hyperparamètre rend son ajustement crucial (Bengio, 2012). L'algorithme de descente du gradient permet d'optimiser les paramètres afin que l'écart entre les sorties obtenues et attendues diminue. Le processus d'apprentissage converge alors vers une solution.

L'algorithme de descente de gradient utilisé est généralement l'algorithme de descente de gradient stochastique (SGD pour *Stochastic Gradient Descent*) (Erickson *et al.*, 2017 ; Lecun *et al.*, 1998). À chaque itération, au lieu de calculer le gradient pour tous les échantillons de l'ensemble de données, l'algorithme SGD calcule le gradient seulement pour un ou quelques échantillons de l'ensemble de données, sélectionnés au hasard. Ce nombre d'échantillons considéré correspond à la taille du lot (*batch size*). L'algorithme SGD permet de réduire considérablement les coûts de calcul et ainsi de réaliser les itérations plus rapidement (Lecun *et al.*, 1998). L'apprentissage s'effectue sur plusieurs époques (*epochs*). Un époque correspond au nombre d'itérations nécessaires pour traiter l'ensemble des données d'entrainement. Ce nombre d'itérations dépend donc de la taille de lot choisie.

Pour la tâche de classification, la fonction de perte d'entropie croisée est communément utilisée (Brownlee, 2020). L'entropie croisée mesure la différence entre la distribution d'un vecteur $p$ de vérité terrain et d'un vecteur de distribution prédit $q$ (Murphy, 2012). Soit $C$ le nombre total de classes à détecter. Le vecteur $p$ se compose des probabilités estimées par le réseau pour chaque classe, comprises entre 0 et 1.

Pour un échantillon donné, la fonction d'entropie croisée (CE pour *Cross Entropy*) s'écrit (Murphy, 2012):

$$CE(p,q) = -\sum_{c=0}^{C-1} p_c \log q_c \tag{1.3}$$

Il existe de nombreuses fonctions de perte utilisées pour résoudre des problèmes de régression, telles que les fonctions de perte L2 et L1. Soient $Y$ le vecteur prédit et $Y'$ le vecteur de valeurs de vérité terrain. Ces vecteurs sont de dimension $(N, 1)$.

La fonction de perte L2, aussi appelée erreur quadratique moyenne (MSE pour *Mean Squared Error*) minimise l'écart au carré entre $Y$ et $Y'$ (Murphy, 2012). Pour un échantillon, L2 s'écrit :

$$L2(Y', Y) = \frac{1}{N}\sum_{n=0}^{N-1} (y'_n - y_n)^2 \tag{1.4}$$

La fonction de perte L2 étant sensible aux valeurs aberrantes, la fonction de perte L1, plus robuste à ces dernières, est parfois préférée (Murphy, 2012). La fonction de perte L1, aussi appelée erreur absolue moyenne (MAE pour *Mean Absolute Error*) minimise l'écart absolu entre $Y$ et $Y'$ (Murphy, 2012). Pour un échantillon, L1 s'écrit :

$$L1(Y', Y) = \frac{1}{N}\sum_{n=0}^{N-1} |y'_n - y_n| \tag{1.5}$$

Les pertes sont en général moyennées sur l'ensemble des échantillons.

### 1.1.3 Les réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNNs) sont des réseaux de neurones artificiels dont l'architecture est spécifiquement conçue pour utiliser des images en entrée. Ils se composent principalement de couches de convolution, de couches de *pooling* et de couches entièrement connectées (Karpathy et Grebu, 2021).

- **Couches de convolution**

Les couches de convolution constituent l'élément fondamental de l'architecture du CNN, elles sont capables d'extraire des caractéristiques depuis des images (Karpathy et Grebu, 2021). Une couche de convolution est composée d'un ensemble de filtres, aussi appelés noyaux de convolution. Une couche de convolution a donc une largeur et une hauteur – dimension des filtres – ainsi qu'une profondeur – nombre de filtres. La dimension des filtres est inférieure à celle de l'image d'entrée sur laquelle la convolution est appliquée.

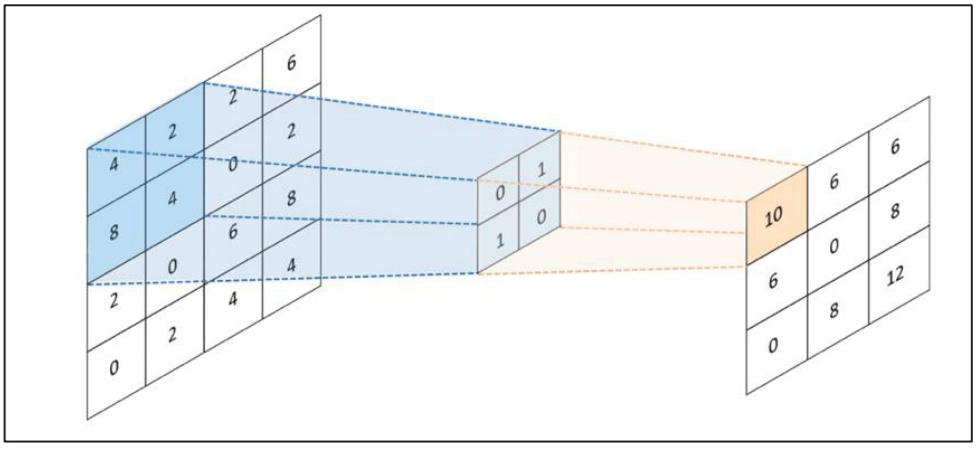


Figure 1.3 Schéma de principe d'une opération de convolution avec un unique noyau et une entrée 2D
Tirée de Tsai, Chen et Wang (2018, p. 5) sous la licence publique Creative Commons Attribution 4.0 Public License

L'opération de convolution s'effectue de la manière suivante : le volume d'entrée est parcouru, à travers toute sa hauteur et largeur, par chaque noyau de convolution. Un produit scalaire est

effectué entre le noyau et chaque zone du volume d'entrée sur laquelle il se déplace (Karpathy et Grebu, 2021 ; Amidi et Amidi, 2021), comme illustré sur la
Figure 1.3. La zone du volume d'entrée considérée par les noyaux de convolution est le champ récepteur. En sortie de l'opération de convolution, on obtient une carte de caractéristiques qui contient la réponse de convolution à chaque position spatiale. Les couches de convolution sont généralement suivies d'une opération de correction ReLU. L'application de la fonction d'activation non linéaire rend possible l'extraction de caractéristiques complexes non linéairement modélisables (Karpathy et Grebu, 2021).

Les hyperparamètres d'une couche de convolution sont la dimension des noyaux de convolution, le pas et le *zero padding* (Amidi et Amidi, 2021). Ils contrôlent la dimension de la carte de caractéristiques en sortie. Le pas correspond au chevauchement entre les champs récepteurs. Le *zero padding* consiste à l'ajout de zéros en bordure du volume d'entrée afin de maitriser les dimensions de la carte de caractéristiques.

Un grand intérêt de la couche de convolution réside dans le fait qu'elle est non entièrement connectée (Karpathy et Grebu, 2021). Le nombre de paramètres à apprendre, correspondant aux poids d'entrée de chaque neurone composant la couche de convolution, est radicalement réduit par rapport à une couche entièrement connectée (Karpathy et Grebu, 2021). Par sa nature, l'opération de convolution est efficace et invariante par translation.

- **Couches de *pooling***

La couche de *pooling* suit généralement le bloc de couche de convolution avec ReLU (Karpathy et Grebu, 2021 ; Murphy, 2012). Elle réalise une opération de sous-échantillonnage : elle sert à réduire le nombre de paramètres à apprendre en réduisant la dimension des cartes d'activation. L'opération de *pooling* est généralement réalisée par un filtre de taille 2x2 et avec un pas de 2 (Karpathy et Grebu, 2021). Parmi les différents types de *pooling*, le *max pooling* est le plus utilisé (Murphy, 2012). Il sélectionne la valeur maximale dans des régions de taille 2x2 de la carte d'activation (voir Figure 1.4).

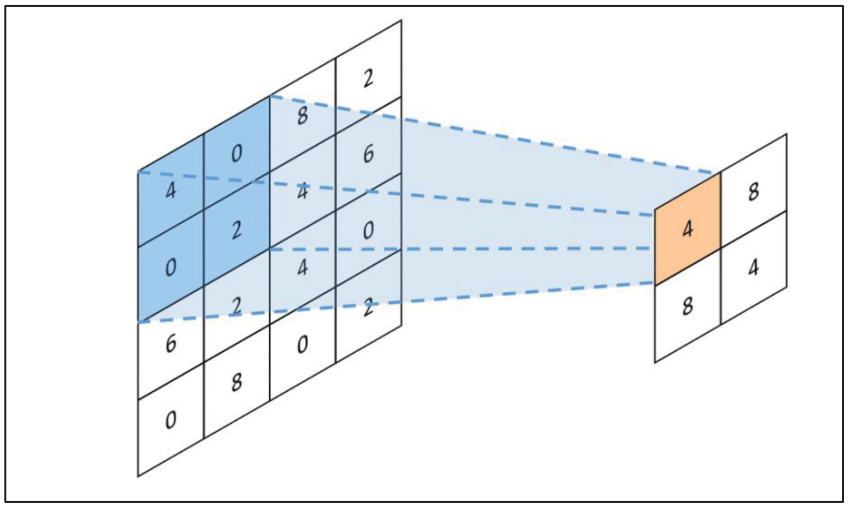



Figure 1.4 Schéma de principe d'une opération de *max pooling*
Tirée de Tsai, Chen et Wang (2018, p. 6) sous la licence publique Creative Commons Attribution 4.0 Public License

- **Couches entièrement connectées**

Les couches entièrement connectées (FC pour *Fully Connected*) se trouvent généralement à la suite de couches de convolution avec ReLU et de *pooling* (Karpathy et Grebu, 2021). Chaque entrée est connectée à chacun des neurones composant la couche FC, permettant ainsi d'apprendre les relations possiblement non linéaires qui lient les caractéristiques extraites dans les couches précédentes (jamesmf, 2017).

### 1.1.4 Les réseaux de neurones profonds

Depuis les années 2010, les réseaux de neurones ont connu d'énormes progrès grâce à la mise en place de méthodes d'apprentissage profond (Zavalina *et al.*, 2020).

L'intuition derrière l'apprentissage profond vient du constat qu'il existe une hiérarchie dans les caractéristiques apprises : le niveau d'abstraction de leur représentation grandit avec le nombre de couches. Le réseau profond, composé de nombreuses couches cachées, montre une très grande capacité de représentation et permet ainsi la modélisation de problèmes complexes (Zavalina *et al.*, 2020). LeCun définit un réseau neuronal profond comme un réseau qui

possède plusieurs couches et qui les utilise afin de construire une hiérarchie des caractéristiques de complexité croissante (Zavalina *et al.*, 2020).

Des réseaux de classification de plus en plus profonds et de plus en plus performants sont développés : le réseau AlexNet (Krizhevsky, Sutskever et Hinton, 2012), fort de 12 couches; l'architecture VGG (Simonyan et Zisserman, 2014), se déclinant en 2 versions fortes de 16 et 19 couches de poids et ResNet (He, X. Zhang, *et al.*, 2016), existant sous les versions ResNet34, ResNet50 et ResNet101, respectivement fortes de 34, 50 et 101 couches de poids. À partir des architectures profondes, de nombreux réseaux de pointe sont développés. Ces derniers sont capables d'effectuer des tâches de détection d'objet et de segmentation. Les CNNs tels que Mask R-CNN (He *et al.*, 2017) et YOLO (Redmon *et al.*, 2016) connaissent depuis un succès incroyable.

Multiples réseaux de détection ou de segmentation, tels que RetinaMask (Fu, Shvets et Berg, 2019) et Mask R-CNN (He *et al.*, 2017) utilisent l'architecture ResNet (He, X. Zhang, *et al.*, 2016), en tant qu'extracteur de caractéristiques, dit *backbone*. ResNet, diminutif de *Residual Network*, implémente l'apprentissage résiduel dans le but de faciliter l'apprentissage dans les couches profondes. Spécifiquement, ResNet vient en solution à un frein majeur à la mise en place de réseaux profonds, le problème de disparition du gradient : le gradient rétropropagé dans le réseau devient très petit, empêchant le changement de valeurs des poids et paralysant ainsi le processus d'apprentissage (He, X. Zhang, *et al.*, 2016).

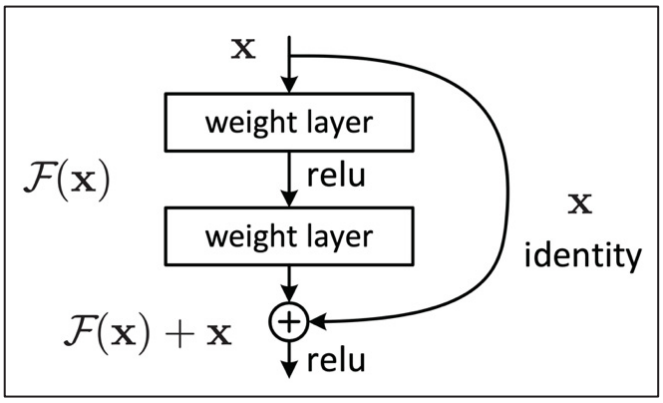



Figure 1.5 Schéma de principe d'un bloc résiduel
Tirée de He, X. Zhang, *et al.* (2016, p. 2) © 2016 IEEE

L'apprentissage résiduel proposé par ResNet (He, X. Zhang, *et al.*, 2016) se fait par l'ajout de connexions sautes-couches (ou résiduelles) entre différentes couches de poids, qui contournent ces couches de traitement, notamment des couches non linéaires. Un bloc résiduel est illustré en Figure 1.5. Dans un bloc résiduel, le réseau apprend la fonction résiduelle $F(x)$, correspondant à la différence entre l'entrée et la sortie (soit le résidu). Grâce aux connexions sautes-couches, les gradients sont rétropropagés dans les couches du réseau sans subir d'importantes diminutions de valeur. En plus de régler le problème de disparition de gradient, l'architecture ResNet est plus facile à optimiser que les architectures sans connexions résiduelles (He, X. Zhang, *et al.*, 2016).

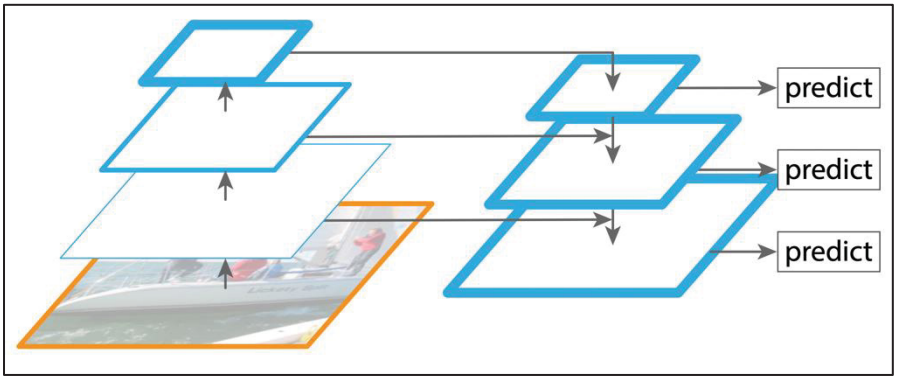



Figure 1.6 Schéma du *Feature Pyramid Network*
Modifiée de Lin, Dollár, *et al.* (2017, p. 2) © 2017 IEEE

Le calcul d'une hiérarchie des caractéristiques faite par un réseau profond tel que ResNet lui impose une structure pyramidale : plus on monte le long de la pyramide, plus les cartes de caractéristiques sont de faible résolution spatiale, mais possèdent des caractéristiques de haut niveau (Lin, Dollár, *et al.*, 2017). La structure FPN (*Feature Pyramid Networks*) (Lin, Dollár, *et al.*, 2017), avec un faible coût de calcul, met à profit la structure pyramidale du CNN en s'assurant que les cartes de caractéristiques possèdent des informations sémantiques solides sur tous les niveaux. Le FPN se compose d'une voie ascendante et d'une voie descendante presque symétriques. Un schéma de principe est donné dans la Figure 1.6. La voie ascendante correspond au *backbone* utilisé, par exemple ResNet. Sur le chemin descendant, les cartes de caractéristiques sont obtenues en joignant la carte de niveau supérieur suréchantillonnée et la

carte de caractéristiques de niveau correspondant sur la voie ascendante. La carte créée contient ainsi des informations sémantiques de haut niveau -grâce à la première carte- et des informations de localisation précises -grâce à la deuxième carte. Les prédictions sont ainsi faites à partir des cartes de caractéristiques de différents niveaux afin de bien détecter les petits comme les grands objets dans l'image (Lin, Dollár, *et al.*, 2017). Le FPN est aujourd'hui adopté par un grand nombre de réseaux de détection et de segmentation (He *et al.*, 2018 ; K. Chen, Pang, *et al.*, 2019 ; Huang *et al.*, 2019 ; Bolya *et al.*, 2019 ; Fu, Shvets et Berg, 2019).

### 1.1.5 Pratiques communes pour l'entrainement supervisé des CNNs

#### 1.1.5.1 Apprentissage par transfert et *fine-tuning*

L'apprentissage par transfert est une technique largement utilisée lors de l'utilisation de CNNs. Elle est censée accélérer la convergence du CNN et booster ses performances (He, Girshick et Dollar, 2019). Elle repose sur le fait que la représentation des caractéristiques apprise par un CNN sur un ensemble de données comporte des informations utiles à l'apprentissage d'une autre tâche (He, Girshick et Dollar, 2019). En pratique, l'apprentissage par transfert consiste à entrainer un modèle CNN sur un ensemble de données source et à utiliser une partie ou l'ensemble du modèle appris comme nouveau modèle à entrainer sur un ensemble de données cible (Ekkis, 2018). Ce nouveau modèle est appelé modèle pré-entrainé.

Le modèle pré-entrainé est alors entrainé sur l'ensemble de données cible selon différentes configurations d'hyperparamètres afin de déterminer celle qui offre les meilleures performances. Cette recherche des hyperparamètres adéquats s'appelle l'ajustement (*fine-tuning*) du système (Brownlee, 2019a).

#### 1.1.5.2 Stratégie de répartition des données en apprentissage supervisé

L'apprentissage d'un CNN requiert des données distinctes pour l'entrainement du modèle et pour son évaluation.

Lorsqu'on applique simplement un modèle à un ensemble de données, ce dernier est séparé, généralement aléatoirement, en un sous-ensemble d'entrainement et un sous-ensemble de test. En revanche, lorsqu'on souhaite modifier et ajuster un modèle pour une tâche, il est nécessaire de séparer les données disponibles selon trois sous-ensembles : l'ensemble d'entrainement, l'ensemble de validation et l'ensemble de test (Brownlee, 2017). Pour chacune des configurations d'hyperparamètres à essayer, le modèle est entrainé sur l'ensemble d'entrainement puis évalué sur le jeu de données de validation. Quand le modèle est ajusté, ce dernier est évalué sur les données test. L'évaluation sur le jeu de données test renseigne sur les performances du modèle sur un ensemble de données qui lui est inconnu et indique donc sa capacité à généraliser. Effectuer l'ajustement du réseau sur un ensemble de validation et non pas directement sur l'ensemble de test permet de prévenir le phénomène de surapprentissage, qui correspond à un modèle trop ajusté à un ensemble de données et qui est incapable de généraliser à de nouvelles données (Brownlee, 2017).

Pour effectuer le *fine-tuning* d'un modèle, la stratégie de répartition des données selon la validation croisée en *k* plis est recommandée (scikit-learn developers, 2021 ; Pramoditha, 2020). La validation croisée en *k* plis (*k-fold cross validation*) aide à mieux estimer les véritables performances d'un modèle tout en profitant de plus de données pour l'entrainement, notamment dans le cas de petits ensembles de données (scikit-learn developers, 2021). Le principe, illustré en Figure 1.7, est de disposer de plusieurs plis d'entrainement et de validation. L'ensemble de données est séparé en un sous-ensemble d'entrainement et un sous-ensemble de test. L'ensemble de test est mis de côté tandis que l'ensemble d'entrainement est divisé en *k* plis. À partir des *k* plis, *k* répartitions distinctes des données sont faites : dans une répartition, *k*-1 plis sont utilisés pour l'ensemble d'entrainement et le pli restant constitue l'ensemble de validation. Sur la Figure 1.7, le pli de validation est représenté en bleu foncé tandis que les plis constituant l'ensemble d'entrainement sont représentés en bleu clair.

Pour effectuer le *fine-tuning*, de multiples modèles, présentant chacun une configuration d'hyperparamètres différente, sont mis en place. Chaque modèle est alors entrainé et validé sur les *k* plis. Ses performances moyennes correspondent à la moyenne de ses performances sur les *k* plis de validation distincts. La stratégie de validation croisée en *k* plis permet de comparer

les résultats d'un même modèle sur différents ensembles d'entrainement et d'avoir une idée de la capacité de généralisation et de la stabilité d'un modèle (Pramoditha, 2020). À partir des résultats moyens des multiples modèles, la meilleure configuration d'hyperparamètres est sélectionnée. Le modèle selon la configuration optimale est alors entrainé sur l'ensemble d'entrainement entier et est évalué sur l'ensemble de test. Les performances sur l'ensemble de test révèlent les réelles performances du modèle, sur un ensemble qu'il n'a jamais vu. On évalue ainsi la capacité de généralisation du modèle (Pramoditha, 2020).

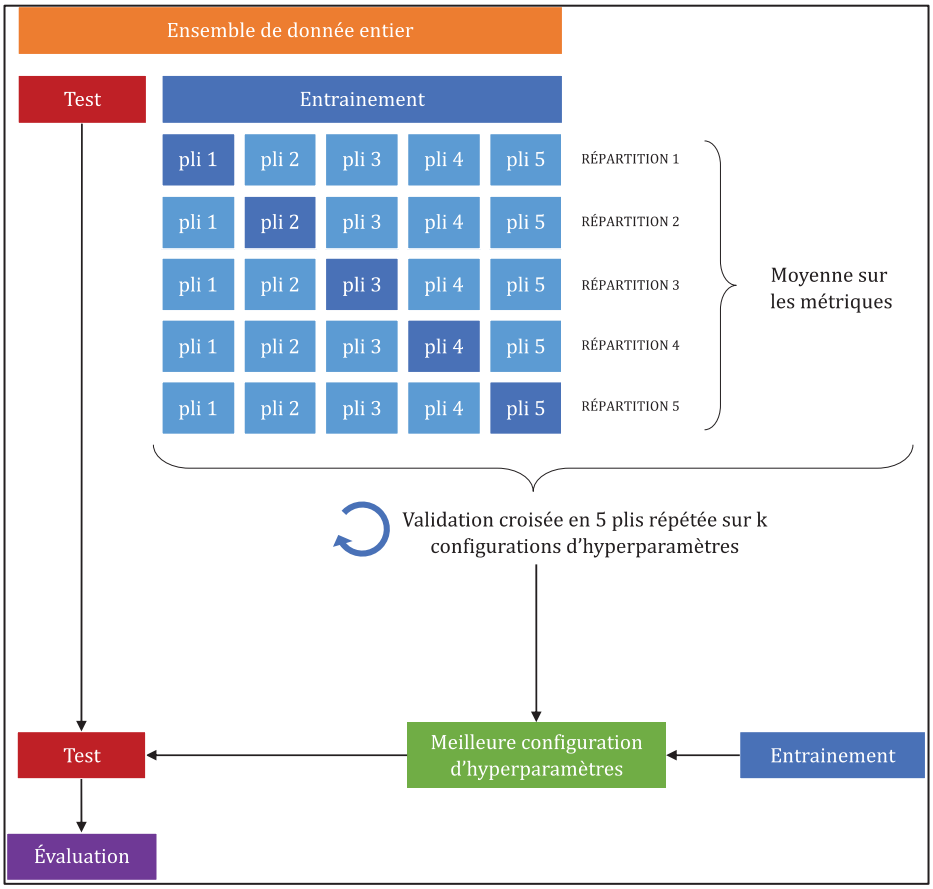



Figure 1.7 Principe de la validation croisée à 5 plis pour le *fine-tuning* d'un CNN
Modifiée de Pramoditha (2020)

La validation croisée implique cependant un coût de calcul additionnel. Les valeurs standard de *k* sont 5 ou 10 (scikit-learn developers, 2021).

#### 1.1.5.3 Diagnostic de l'apprentissage

L'analyse des courbes de perte en entrainement et en validation permet de connaître le comportement du modèle au long de l'entrainement. Cette analyse permet de déterminer si le modèle effectue un sous-entrainement (*underfitting*), un surentrainement (*overfitting*) ou un bon entrainement (*good fitting*) (Brownlee, 2019b).

Un bon entrainement correspond à un modèle ajusté aux données d'entrainement et de validation. Le graphique A. de la Figure 1.8 montre un très bon entrainement. Les courbes de perte en entrainement et en validation diminuent au cours de l'entrainement jusqu'à atteindre un point de stabilité. À partir de ce point, les courbes de validation et d'entrainement montrent un écart réduit.

Un modèle surentrainé correspond à un modèle trop ajusté aux données d'entrainement, qui ne parvient plus à généraliser aux données de validation. Il est représenté sur le graphique B. de la Figure 1.8. Sur les courbes de perte en entrainement et en validation, à partir du point de stabilité atteint, un écart grandit entre les courbes de validation et d'entrainement.

Un modèle est sous-entrainé quand il ne parvient pas à apprendre depuis les données d'entrainement. Comme visible sur le graphique C., l'absence de point de stabilité sur la courbe de perte en entrainement témoigne du phénomène de sous-entrainement (Brownlee, 2019b).

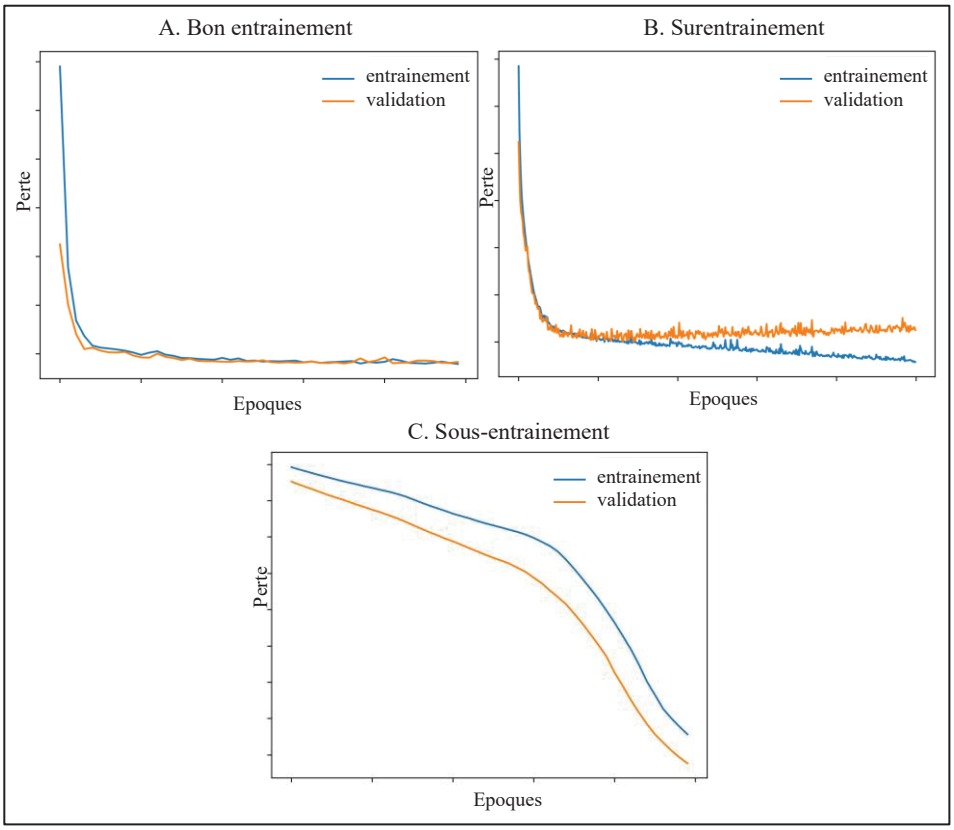



Figure 1.8 Exemples de courbes d'apprentissage pour un bon entrainement (A.), un surentrainement (B.) et un sous-entrainement (C.) Adaptée de Brownlee (2019)

## 1.2 Segmentation et identification d'objets sur des images par CNNs

Il existe différentes structures de CNNs adaptées à différentes tâches. Ce projet vise à mettre à profit les CNNs pour segmenter et identifier individuellement les vertèbres sur une image. De ce fait, la partie suivante s'intéresse aux tâches relatives à la segmentation et à l'identification d'objets. Le principe des tâches de segmentation sémantique, de détection d'objet et de segmentation d'instance ainsi que leurs métriques et structures CNN associées sont développés ci-dessous.

### 1.2.1 Segmentation sémantique

#### 1.2.1.1 Principe de la tâche de segmentation sémantique

La tâche de segmentation sémantique consiste à attribuer une unique catégorie à chaque pixel de l'image (Asgari Taghanaki *et al.*, 2021 ; Li, Johnson et Yeung, 2017). Les pixels d'une même catégorie forment des ensembles de pixels distinguables sur l'image. La segmentation sémantique peut être binaire – les pixels sont classés selon deux catégories, appelées premier plan et arrière-plan – ou multiclasse (Asgari Taghanaki *et al.*, 2021). Sur la Figure 1.9, l'image du chat est segmentée sémantiquement selon quatre classes : chat, ciel, arbres et gazon. Les regroupements de pixels appartenant à une même catégorie s'appellent des masques de segmentation.

Cependant, la segmentation sémantique, raisonnant sur la classification des pixels seulement, ne supporte pas les instances d'objets (Li, Johnson et Yeung, 2017). Sur la Figure 1.9, les vaches ne sont pas séparées en deux objets distincts, mais forment un même masque de segmentation.

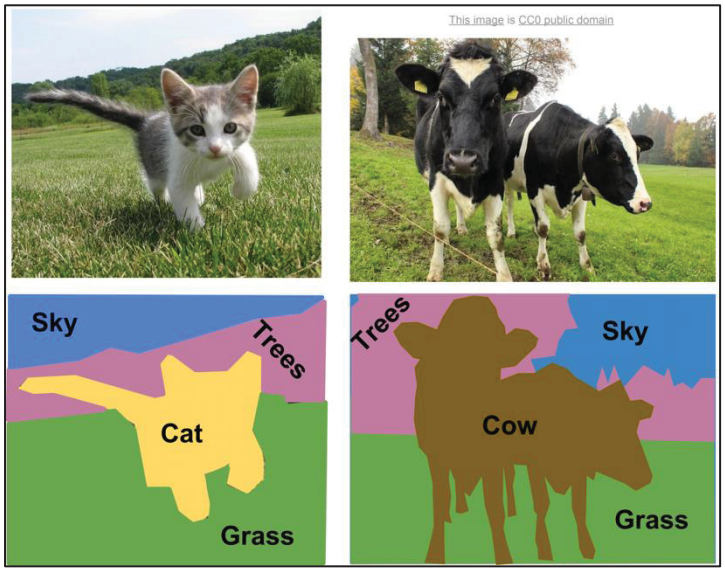



Figure 1.9 Exemples de segmentation sémantique : en haut, les images d'entrée, en bas, les images sémantiquement segmentées
Image reproduite sous licence CC0 public domain

#### 1.2.1.1 Métriques d'évaluation des CNNs de segmentation sémantique

- **L'indice de Jaccard -**

Parmi les différentes métriques mises en place pour évaluer la qualité de la tâche de segmentation, l'indice de Jaccard moyen (MIoU pour *Mean Intersection over Union*) est très populaire (Garcia-Garcia *et al.*, 2017 ; Asgari Taghanaki *et al.*, 2021). La plupart des articles et concours l'utilisent pour donner leurs résultats (Garcia-Garcia *et al.*, 2017).

Soit A et B deux ensembles. L'indice de Jaccard est défini comme le rapport entre l'intersection de A et B et l'union de A et B. Plus l'indice de Jaccard moyen est élevé, plus la segmentation est précise et juste.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1.6}$$

- **Le coefficient de similarité de Dice**

Dans le domaine médical, le coefficient de similarité de Dice (DSC pour Dice *Similarity Coefficient*) est préféré à l'indice de Jaccard (Rizwan I Haque et Neubert, 2020). Soit A et B deux ensembles, la métrique de Dice est définie par l'équation suivante :

$$DSC(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1.7}$$

La principale différence entre ces deux métriques d'évaluation tient dans le fait que l'indice de Jaccard a tendance à pénaliser plus fortement les résultats incorrects que le coefficient de similarité de Dice (Rizwan I Haque et Neubert, 2020). Un schéma de ces métriques est représenté sur la Figure 1.10.

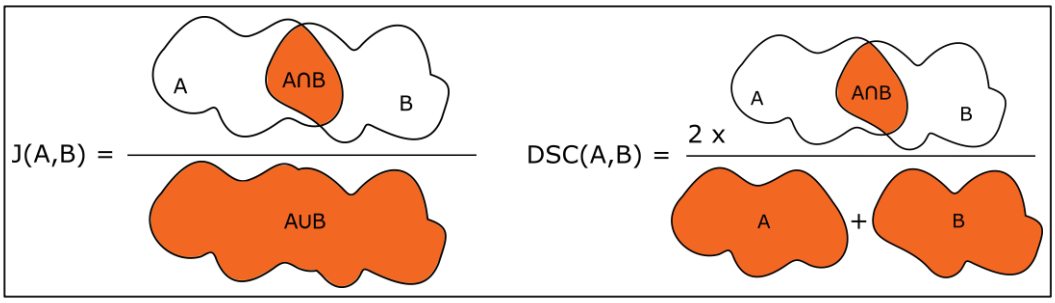



Figure 1.10 Schéma explicatif des métriques du coefficient de Dice et de l'indice de Jaccard

#### 1.2.1.2 U-Net

Le réseau de segmentation sémantique U-Net (Ronneberger, Fischer et Brox, 2015) a été spécialement développé pour la segmentation d'images médicales. Son utilisation étendue aux principales modalités d'imagerie ainsi que l'abondance de papiers proposant des méthodes basées sur l'architecture U-Net témoignent de son immense succès dans le domaine médical (Siddique *et al.*, 2021).

Le principal atout de U-Net est d'être capable de fonctionner avec peu de données annotées tout en produisant des masques de segmentation précis (Ronneberger, Fischer et Brox, 2015). U-Net est un réseau entièrement convolutif (FCN pour *Fully Convolutional Network*) (Long, Shelhamer et Darrell, 2015). Son architecture se divise en deux parties presque symétriques, formant un « U » : le chemin de contraction et le chemin d'expansion. Un schéma de l'architecture de U-net est donné en Figure 1.11.

Le chemin de contraction, à gauche dans la Figure 1.11, suit la structure d'un CNN standard : des blocs constitués de couches de convolution consécutives suivies d'une couche de rectification ReLU ainsi que d'une couche *max pooling* pour sous-échantillonner. Dans chaque bloc du chemin d'expansion, à droite dans la Figure 1.11, une *up-convolution* est effectuée. Elle consiste en une opération de suréchantillonnage suivie d'une opération de convolution. La carte de caractéristiques résultante est ensuite concaténée à la carte de caractéristiques correspondante du chemin de contraction, qui a été préalablement recadrée. La concaténation permet de récupérer, depuis le chemin de contraction, les informations de localisation perdues.

Des opérations de convolution successives suivies d'une opération de rectification ReLU sont enfin réalisées pour obtenir une sortie précise à partir de la carte de caractéristiques concaténée. La dernière couche consiste en une convolution permettant de faire correspondre le vecteur de caractéristiques au nombre de classes désiré (Ronneberger, Fischer et Brox, 2015). L'architecture en « U » présente l'avantage de faire circuler l'information de contexte le long du réseau. La segmentation est ainsi réalisée en connaissance du contexte de la région alentour de l'objet à segmenter.

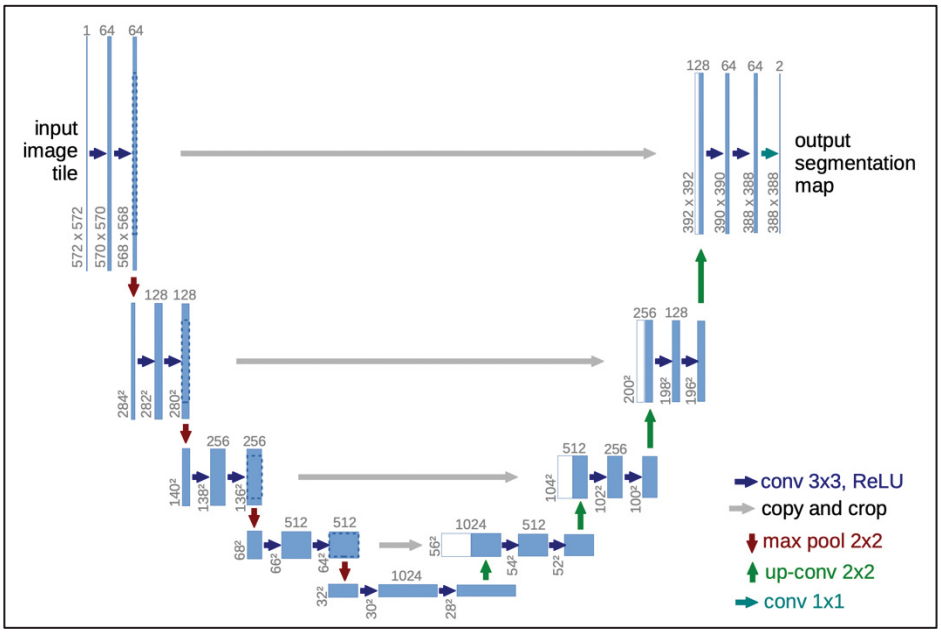



Figure 1.11 Architecture de U-Net
Tirée de Ronneberger, Fischer et Brox (2015, p. 235)
© Springer International Publishing Switzerland 2015

Afin de résoudre l'inconsistance entre le nombre considérable de données annotées nécessaire à l'apprentissage d'un CNN standard et le nombre de données médicales très souvent limité, U-Net repose sur une utilisation approfondie de l'augmentation de données (Ronneberger, Fischer et Brox, 2015). En appliquant des algorithmes de déformation élastique aux données d'entrainement, le réseau est correctement entrainé avec un nombre de données d'entrainement restreint.

### 1.2.2 Détection d'objet

#### 1.2.2.1 Principe de la tâche de détection d'objet

En détection d'objet, deux types d'information sont recherchés : la classification et la localisation d'objets d'intérêt sur une image (Li, Johnson et Yeung, 2017 ; Hafiz et Bhat, 2020). Les objets sont localisés à l'aide de boîtes englobantes (*bounding boxes*) auxquelles une classe et un score de confiance sont attribués. Le score de confiance traduit le niveau de certitude avec lequel le réseau prédit la classe de l'objet (Li, Johnson et Yeung, 2017).

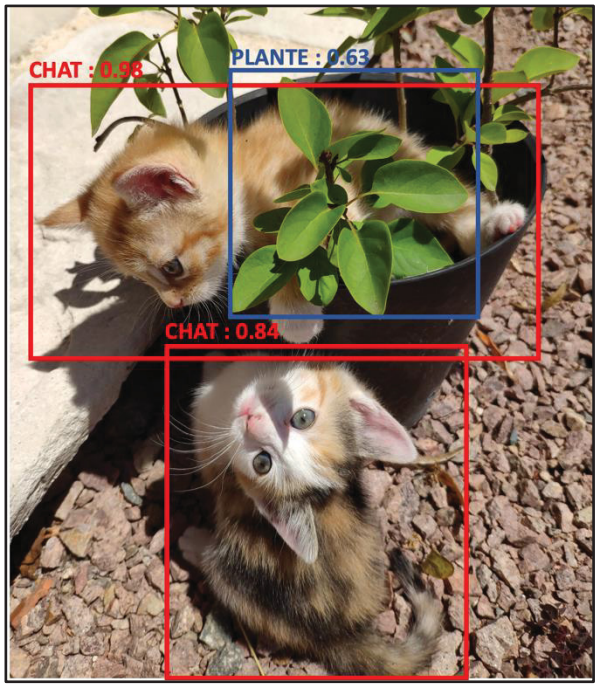



Figure 1.12 Exemple d'une tâche de détection d'objet de deux chats et d'une plante

La tâche de détection d'objet prend en compte les instances : les objets appartenant à une même classe possèdent chacun leur propre boîte englobante, même s'ils se chevauchent (Li, Johnson et Yeung, 2017).

La Figure 1.12 montre un exemple de résultat de détection d'objet : deux chats sont détectés avec des scores de confiance 98% et de 84% ainsi qu'une plante avec un score de confiance de 63%.

#### 1.2.2.2 Métriques d'évaluation des CNNs de détection d'objet

La précision moyenne (AP pour *Average Precision*) est la métrique de référence pour évaluer une tâche de détection d'objet (Zhang et Hong, 2019). Pour comprendre ces métriques, il est nécessaire de connaître les métriques de précision (*precision*) et de rappel (*recall*). La précision et le rappel se calculent à partir des notions de vrai positif (VP), faux positif (FP) et de faux négatif (FN) (Hui, 2018 ; Yohanandan, 2020).

Un CNN de détection d'objet prend en entrée une image et produit en sortie des prédictions, soit des boîtes englobantes avec des labels associés (Li, Johnson et Yeung, 2017). Pour chacune de ces prédictions, l'indice de Jaccard (IoU) est calculé entre la boîte englobante prédite et la boîte englobante de vérité terrain correspondante.

La comparaison entre une valeur seuil d'IoU préalablement fixée et l'IoU relatif à une prédiction permet de désigner cette prédiction comme VP ou FP (Yohanandan, 2020 ; Hui, 2018). Lorsque l'IoU calculé est supérieur au seuil, la prédiction est un VP. À l'inverse, si l'IoU calculé est inférieur au seuil, la prédiction est un FP. Les FN désignent les données de vérité terrain que le CNN n'a pas réussi à prédire (Yohanandan, 2020).

La précision mesure la capacité du modèle à prédire avec exactitude. Elle est définie comme le rapport entre le nombre de VP et le nombre total de positifs prédits.Le rappel mesure la capacité du modèle à trouver les objets dans l'image (Hui, 2018). Il est défini comme le rapport entre le nombre de VP et le nombre total d'objets à détecter dans l'image, soit la somme des VP et FN.

$$Precision = \frac{VP}{VP + FP} \tag{1.8}$$

$$Rappel = \frac{VP}{VP + FN} \tag{1.9}$$

L'AP est définie comme l'aire sous la courbe de précision-rappel. Sa valeur est toujours comprise entre 0 et 1 (Hui, 2018).

Actuellement, la grande majorité des papiers proposant des CNNs de détection d'objet donne leurs performances sur l'ensemble de données très connu COCO (Lin *et al.*, 2014), présentant 330 mille images, 1,5 million d'instances d'objet de 80 catégories différentes. Selon les notations COCO (Lin *et al.*, 2014), l'AP désigne la précision moyenne moyennée sur 10 paliers également espacés de IoU, de 50% à 95%. Les métriques $AP_{50}$ et $AP_{75,}$ désignant respectivement l'AP à des paliers d'IoU de 50% et 75%, sont généralement aussi renseignées.

#### 1.2.2.3 Réseaux de détection d'objet

Il est possible de diviser les CNNs de détection d'objet selon 2 catégories : les CNNs en plusieurs étapes et les CNNs en une étape (Hafiz et Bhat, 2020 ; Lin, Goyal, *et al.*, 2017 ; Soviany et Ionescu, 2018).

- **Réseaux de détection en multiples étapes**

Les réseaux de détection en plusieurs étapes font partie de la famille des réseaux de neurones basés sur les régions (R-CNN pour *Region based Convolutional Neural Network*). Le premier R-CNN est proposé par Girshick en 2016 (Girshick *et al.*, 2016). Plus spécifiquement, les R-CNNs utilisent des algorithmes de prédiction de régions d'intérêt (RoIs pour *Regions of Interest*). Les régions d'intérêt consistent en des régions délimitées de l'image. Ces régions sont ensuite traitées par un réseau détecteur qui génère des prédictions, c'est-à-dire des boîtes englobantes et leurs classes associées. Faster R-CNN (Ren *et al.*, 2016) et Cascade R-CNN sont des réseaux de détection en respectivement deux et quatre étapes.

Faster R-CNN (Ren *et al.*, 2016), amélioration du réseau Fast R-CNN (Girshick, 2015a) qui est lui-même une amélioration de R-CNN (Girshick *et al.*, 2016), est un réseau de détection

très populaire. Faster R-CNN est constitué de deux modules : un réseau de neurones profond entièrement convolutif proposant des régions et un détecteur effectuant les détections à partir des régions proposées.

Un CNN extracteur de caractéristiques, tel que ResNet, est d'abord utilisé. Il correspond au bloc « *conv layers* » sur la Figure 1.13. À partir de la carte de caractéristiques de sortie, un réseau de proposition de régions (RPN pour *Regions Proposal Network*) génère des propositions, notées « *proposals* » sur la Figure 1.13. Le RPN repose sur l'utilisation d'ancres (*anchor boxes*), boîtes de référence suivant plusieurs échelles et formats. Par défaut, Faster R-CNN utilise 9 ancres, correspondant à 3 échelles et 3 formats différents. En se déplaçant sur l'image à la manière d'une fenêtre glissante, des propositions de régions sont générées avec les ancres pour chaque localisation. Une couche de régression et une couche de classification prédisent respectivement les coordonnées des boîtes des régions proposées et un score d'estimation de présence ou non d'objet dans la région. Les propositions redondantes sont supprimées par NMS (*Non-Maximum Suppression*) (Ren *et al.*, 2016). Les RoIs sont ensuite rassemblées dans une nouvelle carte de caractéristiques grâce à une couche de *pooling*, RoIPool. Comme visible sur la Figure 1.13, cette couche RoIPool prend en entrée la carte de caractéristiques résultante du CNN extracteur et les boîtes englobantes des propositions du RPN.

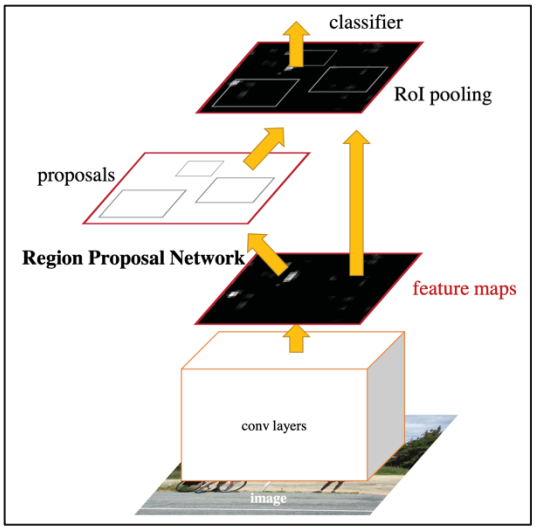



Figure 1.13 Schéma de principe
de Faster R-CNN
Tirée de Ren *et al.* (2016, p. 1139)
© 2016 IEEE

Dans une deuxième étape, le détecteur, aussi appelé tête de prédiction, génère les prédictions à partir des RoIs : composé de couches FC, il effectue la régression des boîtes englobantes et classifie les prédictions (Ren *et al.*, 2016). Les prédictions ont ainsi leur position et format définitif, ainsi qu'une classe attribuée.

Le réseau de détection Cascade R-CNN (Cai et Vasconcelos, 2017) peut être décrit comme un Faster R-CNN avec une architecture plus étendue. Cascade R-CNN est spécialement conçu pour corriger le phénomène de dégradation des performances des CNNs quand le seuil d'IoU, définissant les prédictions négatives et positives, augmente (Cai et Vasconcelos, 2021). De même que pour Faster R-CNN, Cascade R-CNN est composé d'un CNN d'extraction de caractéristiques.

Sur la Figure 1.14, le bloc *conv* extrait les caractéristiques depuis l'image I. L'architecture de Cascade R-CNN diffère de Faster R-CNN par le fait qu'elle comprend trois détecteurs. Les trois détecteurs, notés H1, H2 et H3 sur la Figure 1.14, sont entrainés selon des paliers d'IoU croissants : 0,5; 0,6 et 0,7. Le but de l'architecture en cascade est de mieux différencier les vrais positifs des faux positifs « proches », soit les boîtes englobantes proches de la vérité, mais

non correctes. Les trois détecteurs sont entrainés séquentiellement : la sortie du détecteur d'une étape est utilisée pour entrainer le détecteur de l'étape suivante. De ce fait, à la deuxième étape de détection, réalisée par le détecteur H2, les boîtes englobantes B1 et les classes C1 sont utilisées pour générer les boîtes B2 et les classes C2. Cette technique d'entrainement permet d'entrainer plus efficacement et surtout d'empêcher le phénomène de surapprentissage.

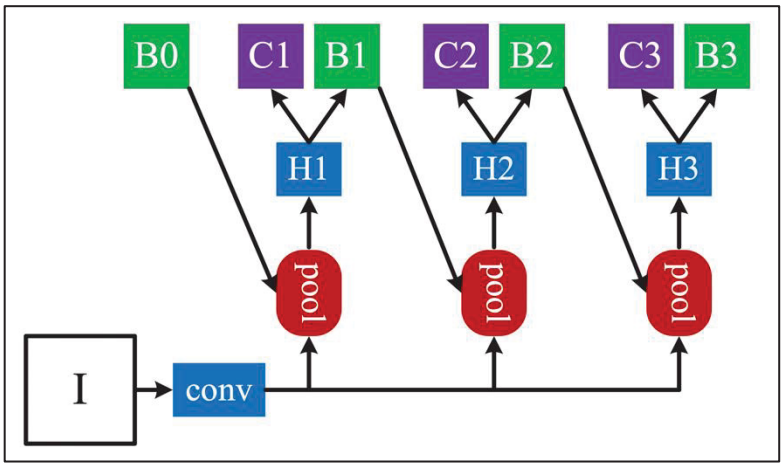



Figure 1.14 Schéma architectural de Cascade R-CNN
Tirée de Cai et Vasconcelos (2021, p. 1486) © 2021 IEEE

- **Réseaux de détection en une étape**

Les CNNs de détection en une étape ne se basent pas sur des régions de l'image, mais sur l'image entière. L'idée est non pas de traiter individuellement chaque région proposée, mais d'envisager la détection d'image comme un problème de régression (Soviany et Ionescu, 2018). Cette particularité rend les méthodes en une étape généralement beaucoup plus rapides, mais moins précises que les méthodes en plusieurs étapes (Hafiz et Bhat, 2020).

RetinaNet (Lin *et al.*, 2018) est un autre réseau de détection d'objet très populaire. Il comprend un ResNet et un FPN en tant que *backbone* puis un module de prédiction. La tête de prédiction, appliquée sur 3 couches de caractéristiques du FPN, est composée d'un sous réseau de régression et d'un sous-réseau de classification (voir Figure 1.15).

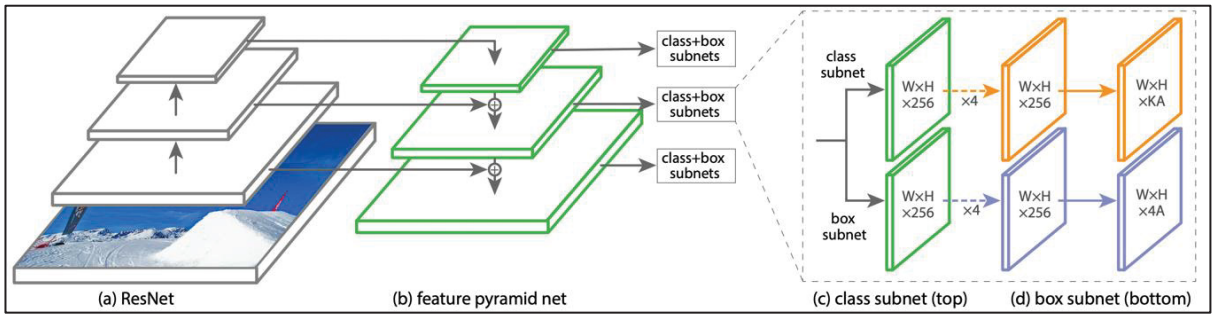



Figure 1.15 Schéma architectural de RetinaNet
Tirée de Lin, Goyal, *et al.* (2017, p. 3003) © 2017 IEEE

La principale motivation des auteurs de RetinaNet est de mettre au point un CNN de détection en une étape aussi précis qu'un CNN en deux étapes en réglant le problème de déséquilibre des classes.

Dans un réseau en une étape, la recherche des objets à différentes localisations, échelles et formats de l'image, va générer plusieurs milliers de propositions (Lin, Goyal, *et al.*, 2017). Parmi ces propositions, seulement une infime partie contient un objet, le reste correspond à l'arrière-plan. Cette surreprésentation de la classe arrière-plan par rapport aux autres classes correspond au phénomène de déséquilibre des classes. Ce déséquilibre est néfaste pour l'apprentissage : la majorité des propositions, classées comme négatives, ne contribuent pas à l'apprentissage qui devient peu efficace ou même incompétent lorsque l'apprentissage est surchargé par les propositions négatives (Lin, Goyal, *et al.*, 2017).

La principale innovation de RetinaNet tient dans sa fonction de perte de classification *Focal Loss* qui résout le problème de déséquilibre de classes. La fonction de perte *Focal Loss* concentre l'entrainement sur les propositions difficiles à classifier afin de diminuer l'influence des propositions faciles à classifier, correspondant majoritairement à la classe arrière-plan.

Soit $y \in \{\pm 1\}$ désignant la vérité de terrain d'une détection. Si la détection appartient à une classe visée, $y = 1$, sinon $y = -1$. Soit $p$ la probabilité estimée par le modèle que la détection appartienne à une classe visée, soit $y = 1$.

On note :

$$p_t = \begin{cases} p & \text{si } y = 1 \\ 1 - p & \text{sinon} \end{cases} \quad (1.10)$$

La fonction *Focal Loss* comprend un terme d'entropie croisée pour la classification binaire et un facteur modulateur $(1 - p_t)$ avec un paramètre modulable $\gamma$.

$$FL(p_t) = -(1 - p_t)^{\gamma} \log (p_t) \quad (1.11)$$

Le paramètre modulable $\gamma$, généralement fixé à 2 permet de régler l'intensité avec laquelle on veut réduire l'influence des propositions faciles à classifier.

### 1.2.3 Segmentation d'instance

#### 1.2.3.1 Principe de la tâche de segmentation d'instance

La tâche de segmentation d'instance peut être vue comme la résolution simultanée de la tâche de détection d'objet et de la tâche de segmentation sémantique (Chen, Hermans, *et al.*, 2017). Cette tâche est complexe, car elle implique la production de résultats à la fois complets et précis en localisation : une boîte englobante contenant un masque de segmentation ainsi qu'un label pour chaque instance d'objet dans l'image.

Par rapport à la détection d'objet, la segmentation d'instance génère une localisation précise avec des masques de segmentation. Par rapport à la segmentation sémantique, elle distingue les différentes instances d'objets appartenant à la même classe (Li, Johnson et Yeung, 2017). La Figure 1.16 montre les résultats selon les trois tâches pour une même image.

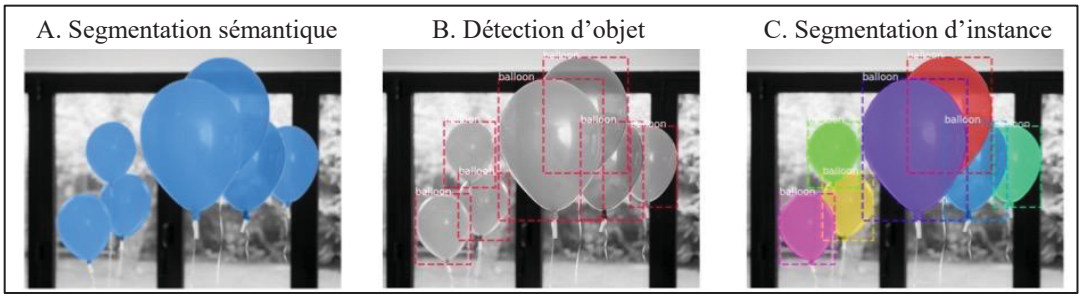



Figure 1.16 Tâche de segmentation sémantique (A.), de détection d'objet (B.) et de segmentation d'instance (C.) appliquées à la même image
Modifiée de Abdulla (2018)

#### 1.2.3.2 Réseaux de segmentation d'instance d'objets

Généralement, les méthodes de segmentation d'instance sont construites en ajoutant une étape de segmentation sémantique à un CNN de détection d'objet (Hafiz et Bhat, 2020) : pour chaque détection faite dans l'image, une tâche de segmentation sémantique est réalisée.
Les réseaux Mask R-CNN, YOLACT, RetinaMask et DetectoRS ont tous été développés à partir de CNNs de détection d'objet.

- **Mask R-CNN**

Mask R-CNN (He *et al.*, 2018), extension de Faster R-CNN (Ren *et al.*, 2016), est un réseau de segmentation d'instance réputé (Hafiz et Bhat, 2020). L'architecture simple de Mask R-CNN repose sur celle de Faster R-CNN, réseau de détection en deux étapes. La première étape comprend un *backbone* ResNet avec un FPN ainsi que le RPN, générant les RoIs. Au niveau de la deuxième étape du réseau, une branche de segmentation est ajoutée parallèlement à la branche de prédiction des classes et des boîtes englobantes. La branche de segmentation est constituée d'un FCN qui effectue la segmentation sémantique binaire de chaque RoI détectée. Mask R-CNN remplace la couche de RoI-*pooling* par la couche RoI-*Align*, qui évite les erreurs d'arrondis sur la localisation spatiale et permet ainsi de générer des masques de segmentation précis (He *et al.*, 2018).

Pour Mask R-CNN et comme pour la majorité des CNNs, le score de confiance d'une prédiction correspond simplement à son plus haut score de classification. La qualité du masque, ne dépendant pas du score de classification, n'est pas reflétée dans le score de confiance. De ce fait, les prédictions avec de hauts scores de confiance, mais des masques médiocres dégradent les performances du CNN. Pour pallier ce problème, MS-RCNN (*Mask Scoring R-CNN*) (Huang *et al.*, 2019) propose un nouveau score de confiance prenant en compte à la fois le score de classification et la prédiction de la qualité du masque. La qualité d'un masque prédit, correspondant à l'IoU entre lui-même et le masque de vérité de terrain correspondant, est notée MaskIoU. Pour prédire le MaskIoU, MS-RCNN ajoute la branche de régression MaskIoU Head à la structure de Mask R-CNN, visible sur le schéma architectural en Figure 1.17. MaskIoU Head, composé de 4 couches de convolutions et de 3 couches complètement connectées, prédit le MaskIoU à partir des caractéristiques des RoIs ainsi que les masques de segmentation correspondants.

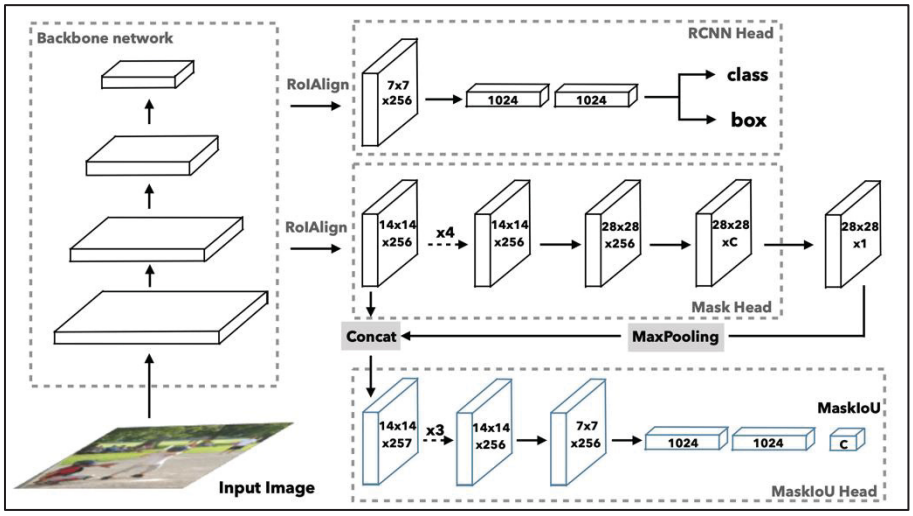



Figure 1.17 Schéma architectural de Mask-Scoring R-CNN
Tirée de Huang *et al.* (2019, p. 6406) © 2019 IEEE

- **RetinaMask**

Le réseau de segmentation d'instance RetinaMask (Fu, Shvets et Berg, 2019) est une amélioration du réseau de détection d'objet RetinaNet. L'architecture de RetinaMask reste très similaire à RetinaNet – un ResNet et un FPN suivis d'une tête de prédiction des classes et des

boîtes englobantes. La principale différence tient dans l'ajout d'un module de prédiction des masques, qui transforme le réseau de détection d'objet en réseau de segmentation d'instance. Les couches de caractéristiques P7, P6, P5, P4 et P3, visibles sur la Figure 1.18, sont utilisées pour prédire les boîtes englobantes et la classe des détections. Les détections, faisant office de propositions de masques, sont assemblées et distribuées sur les couches de caractéristiques P5, P4 et P3. La distribution est gérée par un post-traitement qui détermine la couche adéquate à échantillonner pour prédire le masque de l'instance. P5 permet de prédire les plus gros objets tandis que P3 prédit les plus petits. Chacune des 3 couches de caractéristiques subit une opération RoIAlign avant d'être traitée par un réseau de segmentation FCN.

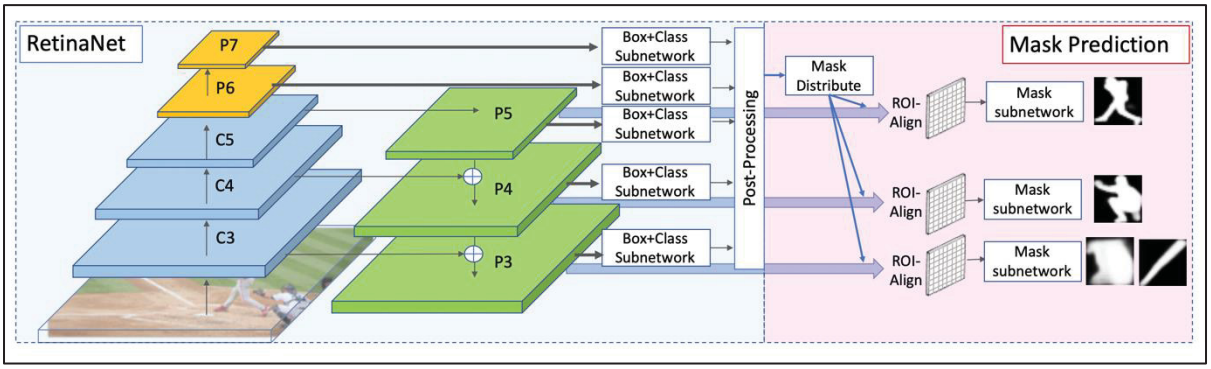



Figure 1.18 Schéma architectural de RetinaMask
Tirée de Fu, Shvets (2019)

- **YOLACT**

YOLACT (You Only Look At CoefficienTs) (Bolya *et al.*, 2019) parvient à exécuter la tâche de segmentation d'instance en temps réel en proposant une méthode en une unique étape, soit en évitant d'avoir recours à une étape de localisation de caractéristiques explicite. La structure de YOLACT est basée sur RetinaNet. Elle consiste en un *backbone* FPN ResNet suivi de deux branches parallèles, une tête de prédiction et le réseau FCN Protonet, dont les sorties sont combinées. Un schéma architectural de YOLACT est représenté sur la Figure 1.19.

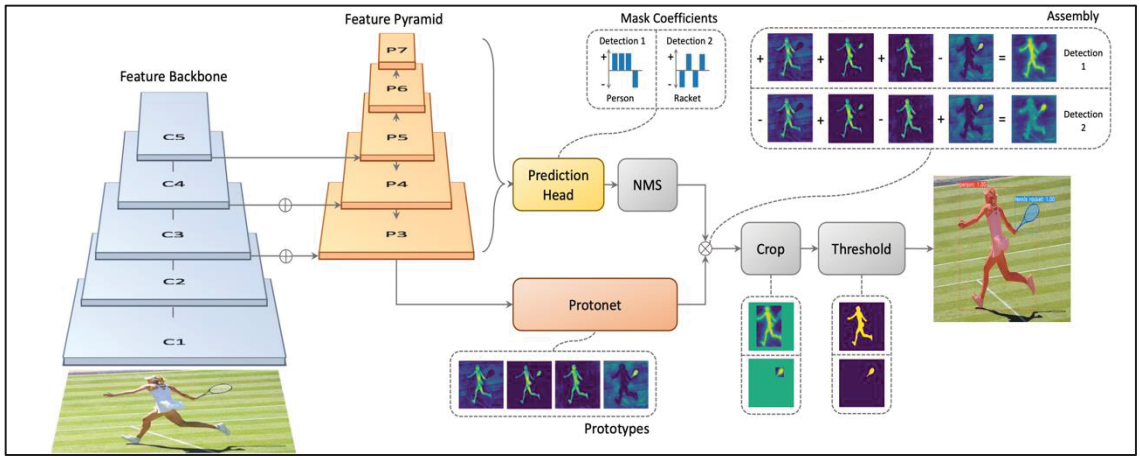



Figure 1.19 Schéma architectural de YOLACT avec $k$ =4
Tirée de Bolya *et al.* (2019, p. 9158) © 2019 IEEE

Afin de se passer de l'étape de localisation de caractéristiques explicite, YOLACT effectue la tâche de segmentation d'instance en la divisant en deux sous-tâches parallèles. La première tâche, gérée par Protonet, consiste en la génération de $k$ images prototypes de masques sans tenir compte des instances. La deuxième tâche est effectuée au sein de la tête de prédiction à laquelle une troisième branche est ajoutée -en plus des branches classification et régression. Cette deuxième tâche correspond en la prédiction de vecteurs de coefficients de masques pour chaque ancre correspondant à une instance. Un vecteur contient $k$ coefficients, soit un pour chaque prototype. Après le filtrage des instances par NMS, les prototypes sont combinés linéairement selon leurs coefficients prédits correspondants. Les masques sont finalement obtenus en recadrant les prototypes selon leur boîte englobante et en appliquant un seuillage.

- **DetectoRS**

Au début de ce projet, DetectoRS (Qiao, Chen et Yuille, 2021) était le réseau le plus performant sur l'ensemble de données COCO test-dev. Il a depuis été détrôné, mais conserve un rang parmi les meilleurs CNNs de segmentation d'instance. DetectoRS (Qiao, Chen et Yuille, 2021) est une amélioration du réseau de segmentation d'instance HTC (*Hybrid Task Cascade*) (Chen *et al.*, 2019), lui-même basé sur le CNN de détection d'objet Cascade R-CNN (Cai et Vasconcelos, 2017).

Le réseau HTC propose d'effectuer la segmentation d'instance en tirant profit de la structure en cascade de Cascade R-CNN. Cependant, le simple ajout de branches de segmentation des masques en parallèle de celles de détection tout en gardant la même architecture n'est pas optimal. En effet, l'exécution parallèle implique que les branches de segmentation reçoivent en entrée seulement des informations relatives aux détections faites à l'étape précédente. Les performances relatives aux masques de segmentation sont ainsi faibles (Chen *et al.*, 2019).
De ce fait, HTC développe une nouvelle architecture maximisant le flux d'informations dans le réseau. La Figure 1.20 illustre l'architecture HTC. « F » représente le *backbone* FPN, « S » la branche de segmentation sémantique, « B » les têtes de prédictions des boîtes englobantes et « M » les têtes de prédiction des masques. Cette structure passe par trois principales améliorations.

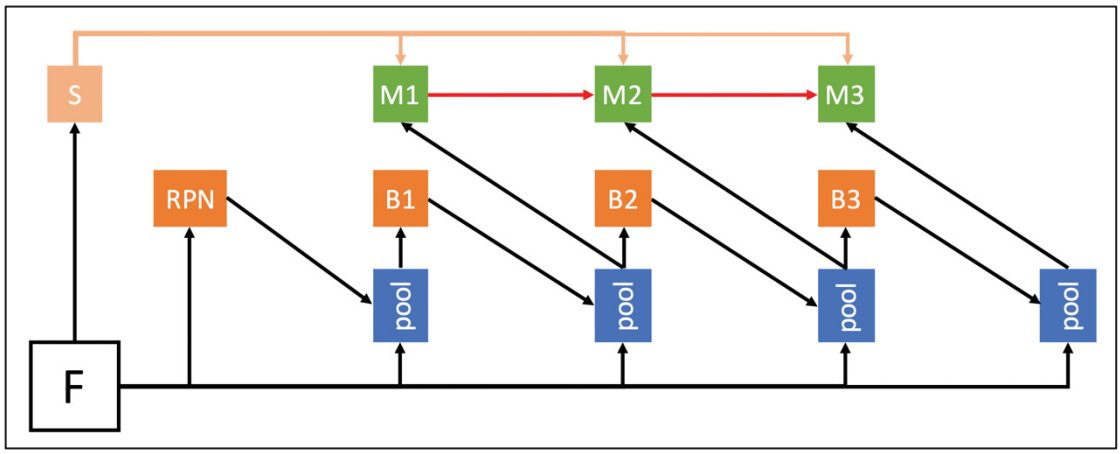



Figure 1.20 Schéma architectural de HTC
Modifiée de Chen K., Pang J., *et al.* (2019, p. 4971) © 2019 IEEE

D'abord, la prédiction des boîtes et des masques d'une même étape n'est pas faite de manière parallèle, mais de manière intercalée. En intercalant les tâches, la tête de prédiction des masques bénéficie des prédictions des boîtes faites à la même étape et génère ainsi de meilleurs masques. Dans le même objectif d'amélioration de l'exactitude des masques, des connexions entre les têtes de prédiction des masques sont aussi ajoutées afin de tirer profit des informations précédemment générées. Elles sont tracées en rouge sur la Figure 1.20. Les masques d'une étape sont donc prédits à partir des caractéristiques extraites par le *backbone*, mais aussi des

masques prédits à l'étape précédente. Ces connexions permettent ainsi l'affinement progressif des masques au cours des étapes. Enfin, afin d'obtenir davantage d'informations relatives au contexte, qui permettent notamment de mieux distinguer les objets de l'arrière-plan, une branche effectuant la segmentation sémantique de l'image est ajoutée. Ces caractéristiques sémantiques extraites sont combinées aux caractéristiques du *backbone* et données en entrée des têtes de prédictions afin d'améliorer leurs résultats. Sur la Figure 1.20, les connexions entre la branche sémantique « S » et les têtes de prédiction des masques « M » sont visibles. Les connexions entre la branche sémantique « S » et les têtes de prédiction des boîtes englobantes n'ont pas été représentées par souci de lisibilité.

La branche de segmentation sémantique est un FCN rattaché sur différentes couches de sortie du FPN afin de capter des caractéristiques de haut, moyen et bas niveau, renseignant des informations à différentes échelles de l'image. Sur la Figure 1.20, les connexions entre la branche sémantique et les branches de détection n'ont pas été représentées par souci de lisibilité.

DetectoRS (Qiao, Chen et Yuille, 2021) propose deux améliorations à HTC en suivant la logique de « voir et penser deux fois » : la structure RFP (*Recursive Pyramid Feature*) et les convolutions à trous commutables (SAC pour *Switchable Atrous Convolution*).

Le RFP se construit à partir d'un FPN, en ajoutant des connexions de rétroaction partant des couches de la voie descendante et se fixant sur les couches de la voie ascendante du *backbone* (Qiao, Chen et Yuille, 2021). En donnant une implémentation séquentielle au RFP, c'est-à-dire en exécutant le *backbone* et le FPN plusieurs fois, le *backbone* analyse et apprend l'image plusieurs fois, suivant les sorties qu'il a générées aux précédentes étapes. Dans l'implémentation par défaut de DetectoRS, le *backbone* et le FPN sont exécutés deux fois (voir Figure 1.21). Le RFP permet à la fois à DetectoRS d'apprendre plus vite et d'atteindre de meilleures performances.

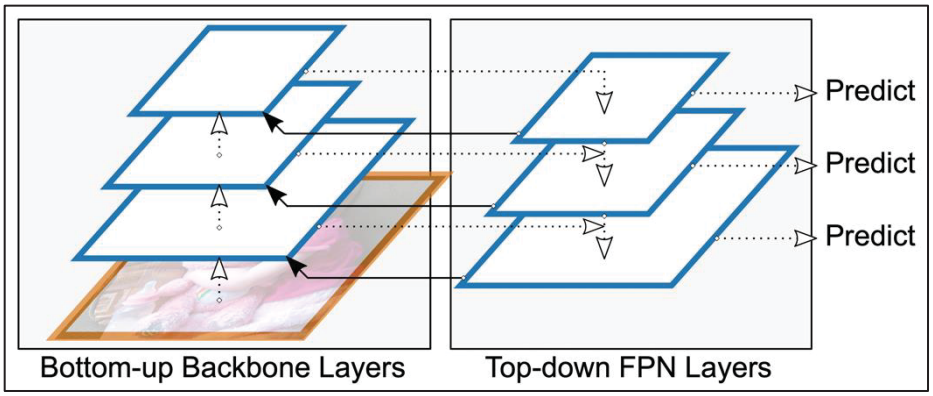



Figure 1.21 Design du Recursive Feature Pyramid
Modifiée de Qiao, Chen et Yuille (2021, p. 10209) © 2021 IEEE

Les convolutions à trous (*atrous convolution*) sont des convolutions comportant un paramètre supplémentaire : le taux à trous. Le taux de trous définit l'espacement entre les valeurs consécutives du noyau de convolution (Pröve, 2017). Un taux de trous $r$ correspond à $r - 1$ zéro entre deux valeurs consécutives du noyau. Sur la Figure 1.22, la convolution à trous en rouge et la convolution à trous en vert montrent respectivement un taux de trous $r = 1$ et $r = 2$. La convolution à trous permet ainsi d'élargir le champ récepteur de la convolution sans impliquer l'apprentissage de paramètres supplémentaires (Pröve, 2017).

La convolution à trous commutable mise en place pour DetectoRS effectue la convolution à trous d'une carte de caractéristiques selon différents taux de trous (Qiao, Chen et Yuille, 2021). Les résultats des convolutions sont ensuite joints avec des fonctions de commutation dépendantes de la carte de caractéristiques en entrée et de la localisation sur cette carte.

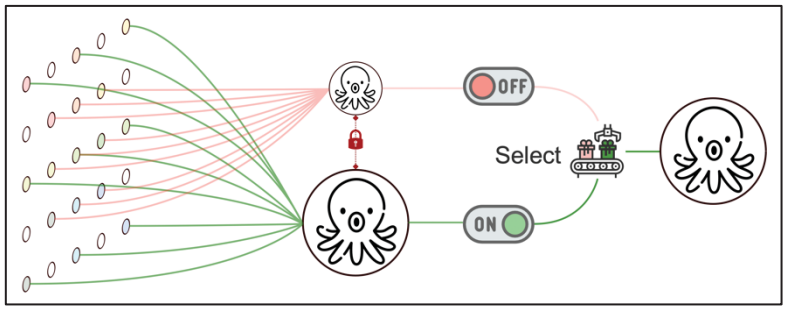



Figure 1.22 Convolution à trous commutable
Modifiée de Qiao, Chen et Yuille (2021, p. 10209)
© 2021 IEEE

### 1.2.4 Conclusion sur les tâches de segmentation et d'identification d'objets

Dans le cadre de ce projet, une simple tâche de segmentation sémantique ne parait envisageable. La tâche de segmentation sémantique, qui attribue une unique classe à un pixel, n'est pas adaptée à la segmentation d'objets superposés tels que les vertèbres sur les radiographies frontales. Les CNNs de détection d'objet, produisant des boîtes englobantes, fournissent des prédictions de localisation trop peu précises. Les méthodes de segmentation d'instance, capable d'identifier et de segmenter des objets même lorsqu'ils sont superposés, semblent bien adaptées au problème.

## 1.3 Segmentation d'instance de vertèbres dans des images radiologiques

Il existe dans la littérature de nombreuses méthodes de segmentation et d'identification de vertèbres sur des images médicales. Alors que ces méthodes étaient généralement basées des outils statistiques et probabilistes (Glocker *et al.*, 2012 ; Rasoulian *et al.*, 2013), les méthodes plus récentes tirent profit des outils d'apprentissage profond, spécifiquement les CNNs.La partie suivante se concentre sur les différentes études qui proposent et testent des méthodes CNNs de segmentation et d'identification des vertèbres sur des images radiologiques.

Bien que la majorité des méthodes CNNs de la littérature s'intéressent soit à la tâche de détection, soit à la tâche de segmentation (Lessmann *et al.*, 2019), certaines méthodes permettent d'effectuer les deux. Les informations d'identification, de localisation, mais aussi de forme des vertèbres sont requises pour certaines applications médicales, notamment pour les méthodes de diagnostic et de planification de chirurgie de la scoliose (Payer *et al.*, 2020).

La localisation et l'identification de vertèbres par réseau de neurones ne sont pas triviales. L'apparence morphologique partagée par les vertèbres rend leur identification exacte complexe (Y. Chen *et al.*, 2019). De plus, les pathologies du rachis peuvent considérablement modifier son apparence globale ainsi que celle des vertèbres. Les variations morphologiques du rachis sont donc nombreuses et plus ou moins intenses (Zhang *et al.*, 2020). Enfin, la superposition des structures anatomiques sur les images radiologiques rend la détection de

repères anatomiques particulièrement compliquée sur ce type d'images médicales (Kim *et al.*, 2021). Les méthodes de segmentation et d'identification de vertèbres sur les images radiologiques peuvent être divisées en deux catégories : les méthodes qui utilisent sans modification des CNNs populaires de segmentation d'instance et les méthodes qui proposent leur propre architecture, spécialement adaptée à la tâche.

### 1.3.1 Application directe de CNNs populaires

Le grand intérêt porté aux méthodes CNNs, leur large potentiel d'amélioration ainsi que les nombreux concours organisés, notamment par COCO, ImageNet et Kaggle Inc., encouragent leur production. Les méthodes CNNs suivent ainsi une constante amélioration. Les CNNs développés au cours des dernières années offrent des performances impressionnantes sur l'ensemble de données COCO (Lin *et al.*, 2014).

De ce fait, certaines études (Wang *et al.*, 2021 ; Yang *et al.*, 2019 ; Kónya *et al.*, 2021) préfèrent appliquer directement des CNNs populaires très performants sur des ensembles de données d'images optiques à leurs propres images médicales plutôt que de concevoir une méthode. Cette partie présente ces travaux ainsi que leurs résultats.

#### 1.3.1.1 Méthodes utilisant des CNNs connus

Dans leurs travaux, (Wang *et al.*, 2021), (Yang *et al.*, 2019) et (Kónya *et al.*, 2021) appliquent des CNNs populaires à la tâche de détection et segmentation de vertèbres sur des images radiologiques. (Wang *et al.*, 2021) et (Yang *et al.*, 2019) emploient le très populaire réseau de segmentation d'instance Mask R-CNN (He *et al.*, 2017). L'étude proposée par (Kónya *et al.*, 2021) compare les réseaux Mask R-CNN (He *et al.*, 2017) , U-Net (Ronneberger, Fischer et Brox, 2015), DeepLabV3 (Chen, Papandreou, *et al.*, 2017), PSPNet (Zhao *et al.*, 2017), et YOLACT.

PSPNet (*Pyramid Scene Parsing Network*) (Zhao *et al.*, 2017) est un FCN intégrant un module de *pooling* en pyramide (*Pyramid Pooling*). Ce module, en effectuant des opérations parallèles

de *pooling* à différentes tailles sur une même carte d'activation, incorpore davantage d'informations de contexte qu'un FCN standard. DeepLabV3 (Chen, Papandreou, *et al.*, 2017) consiste en un FCN incluant le module ASPP (*Atrous Spatial Pyramid Pooling*) permettant de capter les objets, mais aussi des informations de contextes sur différentes échelles de l'image. Le module ASPP effectue, de manière parallèle, des convolutions à trous à différents taux de trous et une opération de *pooling* sur une même carte. Les cartes résultantes sont ensuite concaténées.

- **Données**

(Yang *et al.*, 2019) et (Kónya *et al.*, 2021) utilisent des images radiographiques tandis que (Wang *et al.*, 2021) se sert de données CT-scans (Computed Tomography scans ou scans tomodensitométriques). Plus spécifiquement (Kónya *et al.*, 2021) emploient 730 radiographies latérales des lombaires. (Yang *et al.*, 2019) utilisent des radiographies biplanes du rachis entier acquises avec le système EOS®. Les radiographies en vue frontale et latérale sont jointes sur une unique image de taille fixe. (Wang *et al.*, 2021) s'appuient sur un ensemble de données publique de volumes CT-scans (Computed Tomography scans ou scans tomodensitométriques) dont il extrait des images latérales contenant les vertèbres T1 à L5.
(Yang *et al.*, 2019) et (Wang *et al.*, 2021) prétraitent les données afin de réduire le bruit et les artefacts. Pour cela, (Wang *et al.*, 2021) ajustent l'intensité des images par seuillage et (Yang *et al.*, 2019) utilisent des filtres d'élimination du bruit, médians et d'égalisation d'histogramme. Ils réalisent aussi une augmentation de données pour accroitre la taille des ensembles d'entrainement. (Wang *et al.*, 2021) créent des données supplémentaires, contenant entre 4 et 17 vertèbres, par recadrage des images initiales. (Yang *et al.*, 2019) obtiennent de nouvelles images par rotations aléatoires de ± 3° des données initiales. Le contraste de ces nouvelles images est normalisé.

- **Entrainement**

Les vertèbres à détecter et segmenter varient selon les méthodes. (Wang *et al.*, 2021), s'intéressant aux vertèbres thoraciques et lombaires, attribuent une catégorie différente à

chacune des vertèbres T1 à L5. Les données de vérité terrain sont annotées manuellement. (Yang *et al.*, 2019), qui cherchent à détecter et segmenter la dernière vertèbre cervicale C7 en plus des vertèbres thoraciques et lombaires, ne séparent pas toutes les vertèbres selon différentes classes. En se basant sur un critère de similarité d'apparence, ils distinguent les vertèbres C7, T1 et L5 selon des classes individuelles et attribuent la même classe aux vertèbres T2 à L4 sur la vue frontale. Sur la vue latérale, seules les vertèbres C7 et L5 possèdent leur propre catégorie et les vertèbres T1 à L4 sont rassemblées en une unique catégorie. Les données de vérité terrain sont obtenues par projection des reconstructions 3D du rachis sur les images radiographiques à l'aide de SterEOS©. (Kónya *et al.*, 2021) séparent les vertèbres selon deux classes : les vertèbres lombaires L1 à L5 et la première vertèbre sacrale S1.

De manière très simpliste, (Wang *et al.*, 2021) divisent son ensemble de données en un sous-ensemble d'entrainement de 1260 images et un sous-ensemble de test de 210 images. Mask R-CNN est entrainé puis évalué sur ces derniers. De manière plus élaborée, (Yang *et al.*, 2019) et (Kónya *et al.*, 2021) utilisent la stratégie de validation croisée en 5 plis pour entrainer et évaluer les CNNs.

Dans la méthode proposée par (Yang *et al.*, 2019), les résultats générés par Mask R-CNN sont affinés par un post-traitement avant d'être évalués. Le post-traitement est une fonction polynomiale qui a pour objectif d'écarter les prédictions aberrantes en se basant sur leur position selon l'axe horizontal et d'ajuster la position des centroïdes des vertèbres prédites.

#### 1.3.1.2 Discussion des performances et conclusion

- **Performances de segmentation et d'identification**

Pour Mask R-CNN, (Wang *et al.*, 2021) atteignent un coefficient de Dice moyen de 69,2% et (Yang *et al.*, 2019) un coefficient de Dice moyen pour chacune des catégories sur les vues frontales et latérales toujours supérieur à 87%. L'IoU moyen est supérieur à 80% pour chacun des CNNs dans l'étude menée par (Kónya *et al.*, 2021) et atteint des valeurs maximales de respectivement 85,87% et 87,77% pour Mask R-CNN et YOLACT.

La tâche d'identification est aussi évaluée par (Yang *et al.*, 2019) et (Kónya *et al.*, 2021). En considérant qu'une détection est correcte quand la vertèbre est assignée à la bonne classe et que le score de Dice est supérieur à 50%, (Yang *et al.*, 2019) définissent les métriques de précision et de justesse. La précision correspond au nombre de détections correctes sur le nombre de détections total tandis que la justesse correspond au nombre de vertèbres de réalité de terrain qui sont correctement détectées. Les valeurs de précision et de justesse moyennées sur les catégories sont respectivement de 96,4% et de 98,9% sur la vue frontale. Sur la vue latérale, elles atteignent respectivement 97,7% et 96,8%. (Kónya *et al.*, 2021) utilisent le taux de reconnaissance, qu'il définit de la manière suivante : une image est considérée comme bien reconnue seulement si toutes les vertèbres lombaires et la vertèbre sacrale S1 sont présentes dans des masques distincts. Le taux de reconnaissance ne qualifie pas la qualité des masques de segmentation, il indique seulement la présence de chacune des vertèbres dans des masques distincts.

On remarque que le taux de reconnaissance donné par (Kónya *et al.*, 2021) varie largement en fonction du CNN. Les CNNs Mask R-CNN et YOLACT atteignent des taux de reconnaissance moyens impressionnants de respectivement 99% et 98%. U-Net et DeepLabv3 montrent des taux de reconnaissance de respectivement 91% et 70% alors que PSPNet atteint seulement 43%.

- **Avantages des CNNs de segmentation d'instance sur les CNNs de segmentation sémantique**

Les résultats de (Wang *et al.*, 2021) et (Kónya *et al.*, 2021) tendent à montrer que les modèles de segmentation d'instance performent mieux que les modèles de segmentation sémantique sur la tâche d'identification et de segmentation de vertèbres sur des images respectivement CT et radiographiques. (Wang *et al.*, 2021) entrainent U-Net sur le même ensemble de données que Mask R-CNN. U-Net atteint un coefficient de Dice de seulement 56,9%, soit 12,3 pp (points de pourcentage) en dessous de Mask R-CNN. En plus d'évaluer chacun des cinq CNNs, (Kónya *et al.*, 2021) effectuent une comparaison de leurs performances : des analyses statistiques sont conduites pour mettre en avant les différences significatives entre les modèles.

Ces études statistiques soulignent les écarts de performances entre les deux types de CNNs. Sur la métrique d'IoU moyen, les deux modèles de segmentation d'instance (Mask R-CNN et YOLACT) sont significativement meilleurs que chacun des trois modèles de segmentation sémantique (U-Net, PSPNet et DeepLabV3). Mask R-CNN et YOLACT atteignent des IoU moyens respectifs de 85,87% et de 87,77% contre 83,23% pour DeepLabV3, qui montre le meilleur score sur la métrique parmi les trois modèles de segmentation sémantique. De même, pour le taux de reconnaissance, avec une différence de score de 8 pp, le meilleur modèle d'instance (Mask R-CNN) est significativement meilleur que le modèle sémantique qui performe le mieux (U-Net).

(Wang *et al.*, 2021) et (Kónya *et al.*, 2021) s'accordent aussi à dire que les réseaux de segmentation d'instance, contrairement aux réseaux de segmentation sémantique, présentent l'avantage de pouvoir détecter les vertèbres superposées dans les images latérales. Les réseaux de segmentation sémantique, se basant sur les pixels et non sur les instances, fusionnent les vertèbres se chevauchant dans un unique masque de segmentation. Cette différence de gestion des vertèbres superposées est illustrée en Figure 1.23: alors que la tâche de segmentation d'instance permet de segmenter cinq masques distincts, la segmentation sémantique en distingue seulement quatre, les vertèbres superposées étant réunies dans un unique masque représenté en violet.

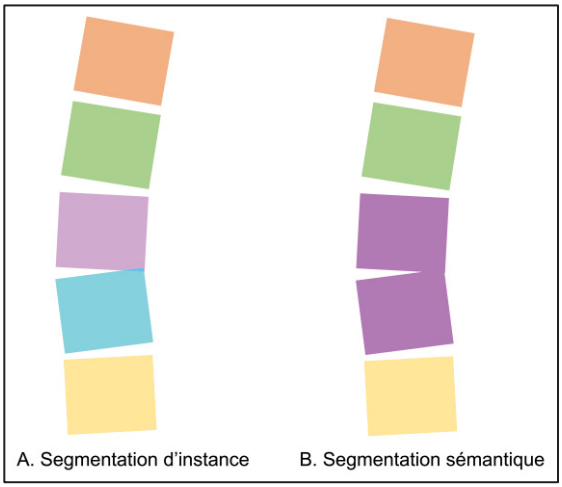



Figure 1.23 Gestion des superpositions par CNN de segmentation d'instance (A.) et de segmentation sémantique (B.)

- **Capacités et faiblesses des CNNs**

Afin d'obtenir des CNNs généralisables à la majorité des patients, (Kónya *et al.*, 2021) construisent un ensemble d'entrainement qui se veut représentatif. Ce dernier inclut de nombreuses images postopératoires contenant des systèmes d'instrumentation tels que des vis et des tiges ainsi que des images de patients atteints de dégénérescence discale. Malgré les difficultés liées à la présence de corps étrangers, de la grande variabilité des espaces intervertébraux et de la superposition accrue des vertèbres, les CNNs sont capables de réaliser la tâche de segmentation avec précision.

(Wang *et al.*, 2021) et (Kónya *et al.*, 2021) remarquent qu'il arrive que les CNNs de segmentation d'instance manquent la prédiction d'une instance ou en prédisent trop. (Kónya *et al.*, 2021) conseillent d'abaisser le seuil de score de prédiction pour régler le problème de sous-prédiction et d'appliquer un algorithme de suppression non maximale (NMS pour *Non Maximum Suppression)* pour la sur-prédiction. Face au problème de sur-prédiction, (Yang *et al.*, 2019) ne considèrent que les détections dont le score de confiance est supérieur à 65% et utilisent le post-traitement, qui retire les prédictions aberrantes. La grande variabilité du coefficient de Dice entre les différentes vertèbres dans l'étude de (Wang *et al.*, 2021) est certainement causée par des prédictions manquées auxquelles un coefficient de Dice de 0% a été attribué. Sur les images extraites de CT de rachis en vue latérale, le coefficient de Dice atteint une valeur maximale de 99,9% pour la lombaire L5 alors qu'il chute à 16,1% pour la vertèbre T3. L'étude de (Wang *et al.*, 2021) semble souffrir d'un réel problème de sous-prédiction qui n'a pas été traité.

(Yang *et al.*, 2019), remarquant que les vertèbres très similaires et rapprochées sont plus difficiles à prédire, simplifient la tâche de classification en réunissant les vertèbres T1 à L4 sur la vue latérale et T2 à L4 sur la vue frontale. De la même manière, le travail de (Kónya *et al.*, 2021) différencie simplement les vertèbres selon deux classes : les lombaires et la sacrale S1. Les résultats donnés par ces études ne traduisent donc pas la classification individuelle des vertèbres, mais une classification groupée des vertèbres.

### 1.3.2 CNNs développés spécifiquement pour la tâche de détection et segmentation des vertèbres

D'autres travaux de segmentation et d'identification de vertèbres sur des images radiologiques ont développé des méthodes CNNs conçues spécialement pour cette tâche. Le concours VerSe (*Large Scale Vertebrae Segmentation Challenge)*, organisé dans le cadre de la conférence internationale MICCAI (*Medical Image Computing and Computer Assisted Intervention*), met en avant nombre de ces méthodes. Le rapport des éditions 2019 et 2020 du concours (Sekuboyina *et al.*, 2021) présente les performances des différents participants ainsi que les méthodes utilisées pour la segmentation et l'identification individuelle de toutes les vertèbres sur des volumes CT-scans du rachis. Des cas de rachis sur- ou sous-numéraires apparaissent dans l'ensemble de données. Les masques de segmentation de vérité terrain sont extraits automatiquement des 374 CT-scans puis sont vérifiés et affinés par différentes équipes d'experts.

(Kim *et al.*, 2021) proposent une méthode spécialement conçue pour la segmentation et l'identification des vertèbres lombaires (L1 à L5) sur des images radiographiques latérales du rachis recadrées sur les lombaires. Les 160 images radiographiques sont annotées manuellement et vérifiées par deux radiologistes.

On retrouve deux points communs dans les différentes méthodes les plus performantes du concours Verse (Sekuboyina *et al.*, 2021 ; Payer *et al.*, 2021 ; Lessmann *et al.*, 2019) et dans celle de (Kim *et al.*, 2021) : toutes ces méthodes mettent en place une stratégie pour effectuer de la segmentation d'instance à partir de réseaux sémantiques basés sur U-Net. De plus, elles intègrent toutes des connaissances préalables.

#### 1.3.2.1 Gestion de la segmentation d'instance

Comme il a été vu précédemment, une simple tâche de segmentation sémantique ne parvient pas à segmenter et identifier les vertèbres de manière optimale sur une image radiologique, notamment à cause des superpositions entre les vertèbres. Les méthodes spécialement développées pour la segmentation et l'identification des vertèbres créent un cadre de

segmentation d'instance dans lequel un variant de U-Net est utilisé pour segmenter les vertèbres. On constate deux stratégies de mise en œuvre de la segmentation d'instance : la mise en place d'un processus itératif et la division de la méthode en plusieurs étapes.

- **Processus itératif**

Dans la méthode qu'ils proposent, (Lessmann *et al.*, 2019) adaptent un U-Net 3D à la tâche de segmentation d'instance en lui faisant suivre un processus itératif. Plus spécifiquement, le processus de détection et de segmentation est guidé pour s'exécuter de manière ordonnée. Le réseau analyse une RoI qui est dirigée dans le volume CT grâce aux informations stockées dans un composant de mémoire. Quand une vertèbre est détectée par le U-Net 3D dans la RoI, cette dernière est déplacée pour se centrer sur cette dernière. La vertèbre est alors segmentée et labellisée par les têtes de segmentation et de classification du U-Net 3D et ces informations sont stockées dans le composant de mémoire. Guidée par les informations données par le composant de mémoire, la RoI se déplace vers la vertèbre suivante. Ce processus guidé par le composant de mémoire permet d'identifier et segmenter chaque instance de vertèbres, une à une, en partant du haut et en descendant dans le volume ou l'inverse. (Lessmann *et al.*, 2019) valident l'utilité de l'approche itérative en entrainant le réseau sans le composant de mémoire.

- **Méthode selon plusieurs étapes**

Les méthodes proposées par (Payer *et al.*, 2020), (Kim *et al.*, 2021) et l'équipe menée par Chen D. dans le concours VerSe (Sekuboyina *et al.*, 2021) se font selon plusieurs étapes. Chaque étape correspond à une tâche et à un CNN pour la résoudre.

(Payer *et al.*, 2020) et Chen D., qui participent aux concours VerSe, suivent une approche de prédiction grossière à fine. (Payer *et al.*, 2020) approximent la localisation du rachis sur le volume CT à basse résolution avec un variant de U-Net. Ensuite, le CNN 3D *SpatialConfiguration-Net* (SCN) (Payer *et al.*, 2019) localise et identifie les centroïdes des vertèbres. L'intérêt de SCN est de mieux localiser les repères anatomiques en apprenant aussi leur configuration spatiale. Cette dernière permet de rendre l'apprentissage plus robuste dans le cas de détections ambigües. Enfin, les RoIs obtenues en sortie de SCN sont données

individuellement en entrée à un U-Net qui effectue une segmentation binaire. Chen D., de la même manière que (Payer *et al.*, 2020), localise grossièrement le rachis sur le volume CT à basse résolution avec un U-Net. Ensuite, à plus haute résolution, un U-Net effectue la segmentation binaire de chacune des vertèbres selon la méthode itérative proposée par (Lessmann *et al.*, 2019). Contrairement à la méthode de (Lessmann *et al.*, 2019), le U-Net itératif ne fait l'identification des vertèbres. Un Resnet50 3D, qui prend en entrée les masques de segmentation obtenus et leur volume CT correspondant, est utilisé pour classifier individuellement les vertèbres.

(Kim *et al.*, 2021) effectuent la segmentation et l'identification des lombaires sur des radiographies en deux étapes. D'abord, les centroïdes des lombaires sont localisés et identifiés à l'aide d'une méthode d'estimation de pose par apprentissage profond, appelée Pose-net (Kim *et al.*, 2021). Pour chaque lombaire détectée, des boîtes englobantes de taille fixe sont calculées à partir des centroïdes. Le CNN M-net (Fu *et al.*, 2018) extrait les masques de segmentation à partir des boîtes englobantes de chaque lombaire, calculés avec les centroïdes détectés. M-Net est un variant de U-Net intégrant des couches multi-échelles et une fonction de coût multi-labels pour mieux apprendre les informations de contexte locales et globales. Les résultats de segmentation sont enfin affinés par une méthode des surfaces de niveau.

#### 1.3.2.2 Mise à profit des connaissances préalables :

L'intérêt de développer une méthode spécialement conçue pour une tâche est de pouvoir l'ajuster à cette tâche. La structure anatomique fixe et rangée de la colonne vertébrale est un élément invariable qui constitue ainsi une source de connaissances préalables très intéressante pour réaliser la tâche de segmentation et de détection. Les méthodes revues mettent à profit ces connaissances préalables par différents moyens.

Le processus itératif suivi par la méthode de (Lessmann *et al.*, 2019) est entièrement basé sur l'information d'arrangement vertical des vertèbres dans la colonne vertébrale. (Payer *et al.*, 2020) utilisent aussi des informations relatives au positionnement des vertèbres dans le rachis. Avant d'effectuer la segmentation par U-Net, les prédictions générées par le SCN sont post-

traitées. Ce traitement sélectionne la séquence de prédictions qui respecte le mieux les contraintes anatomiques d'espacement minimum et maximum entre les vertèbres.
(Kim *et al.*, 2021) s'appuient sur la position particulière de la lombaire L5 pour détecter et identifier les autres lombaires.
L'équipe de Chen D. (Sekuboyina *et al.*, 2021) tire parti de la structure *Deep Reasoning Networks* (DR-Nets) pour contraindre les prédictions des réseaux utilisés dans sa méthode. La méthode de l'équipe de Chen D. ne fait pas l'objet d'un article, mais est seulement très succinctement expliquée dans la revue du concours VerSe. De ce fait, certaines informations quant à l'application DR-Nets dans la méthode manquent. DR-Nets est une structure qui encode les données d'entrée dans un espace latent qui saisit leur structure et les contraintes de connaissances préalables. Un décodeur génère ainsi des sorties qui doivent être cohérentes avec les connaissances préalables. D'après le rapport VerSe, la méthode de l'équipe de Chen D. s'assure grâce à DR-Nets de produire des résultats anatomiquement corrects en termes d'ordre des vertèbres et de masques de segmentation.

#### 1.3.2.3 Discussion des performances et conclusion

Dans le cadre du concours VerSe, le taux d'identification moyen (taux id) et le coefficient de Dice moyen sont calculés pour évaluer les méthodes participantes. Le taux d'identification correspond au rapport entre le nombre de vertèbres bien identifiées et le nombre de vertèbres présentes dans le volume CT. En calculant les distances entre les vertèbres par rapport à leur centroïde, une vertèbre prédite est considérée comme bien identifiée si la vertèbre de vérité terrain la plus proche est celle de catégorie correspondante et si une distance inférieure à 20 mm sépare ces deux vertèbres. Les résultats de (Lessmann *et al.*, 2019), Chen D. et (Payer *et al.*, 2020) qui participent respectivement aux éditions VerSe 2019, 2020 et les deux, sont rapportés dans le Tableau 1.1. Sur l'ensemble de test et sur les trois métriques, (Payer *et al.*, 2020) dominent la compétition pour l'édition 2019 tandis que l'équipe de Chen D. la domine pour l'édition 2020. On remarque aussi que pour les trois méthodes, les écarts de performances

entre l'ensemble de validation et l'ensemble de test sont restreints, témoignant de la bonne généralisation des méthodes.

Tableau 1.1 Résultats du concours VerSe 2019 et 2020 pour les méthodes revues. Tiré de Sekuboyina *et al.* (2021, p. 9)

| | Équipe | Ensemble de validation | | | Ensemble de test | | |
|---|---|---|---|---|---|---|---|
| | | taux id (%) | $d_{moyenne}$ (mm) | Dice (%) | taux id (%) | $d_{moyenne}$ (mm) | Dice (%) |
| Édition 2019 | Payer C. | 95,65 | 4,27 | 90,90 | 94,25 | 4,80 | 89,80 |
| | Lessmann N. | 89,86 | 14,12 | 85,08 | 90,42 | 7,04 | 85,76 |
| Édition 2020 | Chen D. | 95,61 | 1,98 | 91,72 | 96,58 | 1,38 | 91,23 |
| | Payer C. | 95,06 | 2,90 | 88,82 | 92,82 | 2,91 | 89,71 |

Les résultats en hausse pour VerSe 2020 par rapport à VerSe 2019 peuvent en partie s'expliquer par la surreprésentation des cas de rachis surnuméraires (avec des vertèbres T6 et T13) dans l'ensemble d'entrainement VerSe 2020. L'occurrence renforcée de ces cas particuliers permet aux méthodes de mieux les apprendre. On remarque aussi que les méthodes qui prennent en entrée des données 3D atteignent de meilleurs résultats que les méthodes qui prennent en entrée des tranches 2D de volume CT. Cela s'explique par le fait que les informations 3D contiennent beaucoup plus d'informations de contexte, qui jouent un rôle crucial dans la segmentation et l'identification des vertèbres. En revanche, le choix d'une stratégie en une ou plusieurs étapes ne semble pas influencer les résultats, les deux stratégies se retrouvant dans les méthodes les plus performantes.

Pour évaluer sa méthode CNN, (Kim *et al.*, 2021) calculent un taux de succès en sortie du Pose-Net. Un succès correspond au cas où les cinq lombaires sont bien prédites sur l'image, c'est-à-dire que les centroïdes sont au nombre de cinq et localisés sur la vertèbre de bonne catégorie. Le taux de succès sur les 160 radiographies est de 96.25% et une distance moyenne de 5.07 ± 2.17 mm sépare les centroïdes prédits et les centroïdes de vérité. La segmentation des lombaires, calculée après affinement des masques par la méthode des surfaces de niveau,

atteint un coefficient de Dice de 91.60 ± 2.22%. La méthode de (Kim *et al.*, 2021), entrainée et testée sur des images radiographiques latérales, est difficilement comparable aux méthodes VerSe, basées sur des CT-scans.

### 1.3.3 Conclusion sur les méthodes de segmentation d'instance de vertèbres sur des images radiologiques

Le Tableau 1.2 présente un récapitulatif des méthodes discutées dans la revue de littérature. Contrairement aux CNNs de segmentation d'instance, qui sont par définition capable de réaliser la tâche de segmentation d'instance, les méthodes spécialisées, basées sur le CNN de segmentation sémantique U-Net, sont contraintes de mettre en place un cadre de segmentation d'instance en passant par un processus itératif ou à plusieurs étapes. Par rapport à l'utilisation directe de CNNs de segmentation d'instance, les méthodes spécifiquement développées pour l'identification et la segmentation des vertèbres mettent à profit des informations préalables relatives au contexte d'application. Des connaissances relatives à l'anatomie du rachis sont intégrées au sein du processus d'identification et de segmentation permettant à la méthode de produire des résultats plus anatomiquement vraisemblables.

Dans la suite de la revue de littérature, d'autres stratégies de spécialisation d'un CNN à un cadre d'utilisation précis sont présentées.

Tableau 1.2 Résumé des méthodes de segmentation et d'identification des vertèbres revues

| Article | Données d'entrée | Méthode | Gestion de la segmentation d'instance | Intégration savoir préalable supplémentaire | Performances (%) |
|---|---|---|---|---|---|
| (Kónya *et al.*, 2021) | Radiographies latérales des lombaires et de S1 | YOLACT, Mask R-CNN, U-Net, PSPNet, DeepLabV3 | YOLACT, Mask R-CNN : CNNs de segmentation d'instance<br>U-Net, PSPNet, DeepLabV3 : non gérée | Aucune | IoU :<br>Mask R-CNN : 85,87<br>YOLACT : 87,77<br>Taux de reconnaissance:<br>Mask R-CNN : 99<br>YOLACT : 98 |
| (Wang *et al.*, 2021) | Images 2D extraites de CT-scans | Mask R-CNN | CNN de segmentation d'instance | Aucune | Dice : 69,2 |
| (Yang *et al.*, 2019) | Radiographies bi-planes EOS | Mask R-CNN et post- traitement | CNN de segmentation d'instance | Post-traitement basé sur la forme longitudinale du rachis | Frontal / Latéral<br>Dice : 89,8 / 88,7<br>Précision : 96,4 / 97,7<br>Justesse : 98,9 / 96,8 |
| (Lessmann *et al.*, 2019) | CT-scans latéral de VerSe | U-net 3D itératif | Processus itératif qui identifie et segmente les vertèbres de manière ordonnée | Ordre anatomique fixe des vertèbres dans le rachis avec le processus itératif | Dice : 85,76<br>Taux id : 90,42 |
| (Payer *et al.*, 2021) | CT-scans latéral de VerSe | - U-Net : localisation du rachis<br>- SpatialConfiguration-Net : localisation et identification des centroïdes<br>-U-Net : segmentation des vertèbres | Les étapes de localisation et d'identification des centroïdes permettent d'effectuer une segmentation binaire des vertèbres | Apprentissage de la configuration spatiale des vertèbres<br>Choix de la séquence de prédictions respectant des contraintes anatomiques | 2019 / 2020<br>Dice : 89,80 / 89,71<br>Taux id : 94,25 / 92,82 |
| Chen D.(Sekuboyina *et al.*, 2021) | CT-scans latéral de VerSe | - U-Net : localisation du rachis<br>- U-net 3D itératif : localisation et segmentation<br>- ResNet50/DRNets : identification | Les étapes de localisation et d'identification des centroïdes permettent d'effectuer une segmentation binaire des vertèbres | DRNets encode en le savoir à priori et propose des résultats anatomiquement possibles | Dice : 91,23<br>Taux id : 96,58 |
| (Kim *et al.*, 2021) | Radiographies latérales des lombaires | - Pose-net : détection des centroïdes<br>- M-net : segmentation des lombaires<br>- Ajustement des masques | Les étapes de localisation et d'identification des centroïdes permettent d'effectuer une segmentation binaire des vertèbres | Prise en compte de l'ordre anatomique | Dice : 91,60<br>Taux de succès : 96,25 |

## 1.4 Stratégies de spécialisation d'un CNN à un cadre d'utilisation précis

Il a été précédemment montré que spécialiser un CNN à son cadre d'utilisation précis, notamment en tirant profit des connaissances préalables, permet d'améliorer ses performances. Dans la littérature, on retrouve plusieurs solutions pour adapter un CNN à un cadre d'utilisation particulier. Les stratégies de pré-entrainement ajusté et d'ajout de connaissances préalables sous la forme de contraintes sont revues ci-dessous.

### 1.4.1 Apprentissage par transfert ajusté au cadre d'utilisation

#### 1.4.1.1 Application de l'apprentissage par transfert : limites et bénéfices

En imagerie médicale, les *backbones* des nombreux CNNs de classification, détection ou segmentation sont quasiment systématiquement pré-entrainés sur ImageNet (Girshick, 2015b ; Ekkis, 2018). Cependant, plusieurs papiers mettent en doutes les bénéfices apportés par l'apprentissage par transfert. (He, Girshick et Dollar, 2019) avancent que l'apprentissage aide surtout les CNNs à converger plus vite. (Raghu *et al.*, 2019) précisent que l'accélération de la convergence n'a lieu que pour les grands modèles comme ResNet, qui nécessitent un entrainement sur des milliers de données. Le gain sur le temps de convergence est plus marqué quand les représentations des caractéristiques sont transférées seulement sur les couches les plus basses. Les deux articles s'accordent à dire que l'apprentissage par transfert n'affecte pas significativement les performances : elles ne sont ni meilleures ni pires. La généralisation des représentations des caractéristiques d'une tâche à l'autre n'est pas évidente (Kornblith, Shlens et Le, 2019). (Yosinski *et al.*, 2014) démontrent notamment que la transférabilité des caractéristiques s'amoindrit quand les ensembles de données source et cible montrent une grande dissimilitude. (Raghu *et al.*, 2019) rappellent que les images médicales et les images contenues dans ImageNet diffèrent sur de nombreux points : les images médicales sont généralement plus grandes et beaucoup moins nombreuses et les objets d'intérêt qu'elles contiennent, moins discernables, se détectent par des variations très locales.

#### 1.4.1.2 Pré-entrainement sur des images anatomiquement similaires

(Romero *et al.*, 2020) cherchent à établir une méthode adéquate d'apprentissage dans le cas d'entrainement de CNNs sur de petits ensembles d'images radiographiques. Ils démontrent que l'utilisation d'ensembles de données source et cible montrant la même région anatomique rend l'apprentissage par transfert plus efficace. Spécifiquement, les réseaux ResNet50 et DenseNet121 sont entrainés pour la tâche de détection de maladies pulmonaires sur des radiographies de thorax. Ils pré-entrainent ces CNNs sur trois types d'ensembles de données sources : ImageNet, l'ensemble MURA (*Musculoskeletal Radiographs)* de radiographies ciblant différents os du corps et les ensembles de radiographies de thorax librement disponibles CheXpert (Irvin *et al.*, 2019) et Chest X-ray13 (Wang *et al.*, 2017). Par rapport aux pré-entrainements sur MURA et ImageNet qui performent de manière similaire, le pré-entrainement sur les ensembles CheXpert et Chest X-ray13 offre des résultats significativement meilleurs.
(Romero *et al.*, 2020) montrent aussi que le pré-entrainement est plus bénéfique quand il est utilisé comme initialisation des poids que quand il fixe les poids sur la majorité des couches.

### 1.4.2 Ajout de connaissances préalables explicites sous la forme de contraintes

#### 1.4.2.1 Intégration de connaissances préalables explicites dans les CNNs suivant un entrainement peu supervisé

La labellisation des images médicales étant coûteuse en temps et en experts, la recherche tend vers le développement de méthodes CNNs nécessitant moins d'annotations, telles que les méthodes par apprentissage semi-supervisé, par apprentissage non supervisé et par apprentissage faiblement supervisé. Ces dernières performent en général le mieux (Chan, Hosseini et Plataniotis, 2021). Pour la tâche de segmentation par apprentissage faiblement supervisé, les annotations de vérité de terrains correspondent par exemple à des boîtes englobantes, des notes ou des labels (Chan, Hosseini et Plataniotis, 2021). L'objectif est

d'atteindre les performances des apprentissages complètement supervisés avec des apprentissages faiblement supervisés.

Intégrer des connaissances préalables additionnelles de manière explicite dans l'apprentissage est essentiel afin de bénéficier au maximum des informations disponibles. Différents travaux proposent d'imposer au CNN des connaissances préalables sous forme de contraintes (Pathak, Krähenbühl et Darrell, 2015 ; Márquez-Neila, Salzmann et Fua, 2017 ; Jia *et al.*, 2017 ; Zhou *et al.*, 2019 ; Kervadec *et al.*, 2019). Les contraintes peuvent prendre deux types de forme : les contraintes rigides (*hard constraints*) et les contraintes souples (*soft constraints*) (Márquez-Neila, Salzmann et Fua, 2017).

Par rapport aux contraintes souples, les contraintes rigides ont l'avantage d'être toujours respectées. Cependant, les contraintes rigides sont plus difficiles à mettre en place dans une structure d'apprentissage profond : les milliers de paramètres contenus dans un CNN profond rend impossible l'utilisation directe d'une technique d'optimisation contrainte (Pathak, Krähenbühl et Darrell, 2015 ; Márquez-Neila, Salzmann et Fua, 2017).

Pour intégrer des contraintes rigides, (Pathak, Krähenbühl et Darrell, 2015) fusionnent ces dernières aux annotations partielles existantes pour synthétiser des annotations plus complètes. (Márquez-Neila, Salzmann et Fua, 2017) passent eux par l'approche du sous-espace de Krylov pour mettre à profit les contraintes rigides. Il est montré par deux fois que, bien que rien n'impose qu'elles soient toujours respectées, les contraintes souples atteignent de meilleurs résultats que les contraintes rigides (Márquez-Neila, Salzmann et Fua, 2017 ; Kervadec *et al.*, 2019).

Les contraintes souples consistent en des pénalités ajoutées à la fonction de perte. Elles sont simples à mettre en œuvre, mais présentent deux principaux désavantages : il est nécessaire de trouver la pondération appropriée pour chacune des pénalités ajoutées à la fonction de perte et, étant donné leur nature, elles ne sont pas obligatoirement respectées, contrairement aux contraintes rigides.

#### 1.4.2.2 Pénalités anatomiquement contraignantes

Les travaux de (Jia *et al.*, 2017) et (Kervadec *et al.*, 2019) cherchent chacun à améliorer les résultats d'une méthode d'apprentissage peu supervisé en intégrant des contraintes souples. Ces dernières consistent en l'ajout d'un terme de pénalité $l_c$ dans la fonction de perte initiale $l_i$. La fonction de perte totale devient alors $l_t = l_i + l_c$ .

Après avoir remarqué que leur CNN était enclin à surestimer la taille des régions cancéreuses sur les images, (Jia *et al.*, 2017) mettent en place une pénalité basée sur la taille des régions cancéreuses estimée par les experts pour chaque image. Quand une image traitée contient réellement des régions cancéreuses et que la mesure globale des régions détectées comme positives est supérieure à la surface estimée par les experts, une pénalité de type L2 est appliquée dans la fonction de perte. Soit $y$ l'indicateur d'une image cancéreuse, $a$ la surface des régions cancéreuses estimées par les experts et $v$ la surface globale des régions cancéreuses prédites, le terme de pénalité $l_c$ s'exprime selon l'équation (1.12).

$$l_c = \begin{cases} (v-a)^2, & \text{si } y = 1 \text{ et } v > a \\ 0 \quad , & \text{sinon} \end{cases} \tag{1.12}$$

De la même façon, (Kervadec *et al.*, 2019) implémentent une contrainte d'inégalité pour contrôler la taille des masques de segmentation d'organes prédits. À défaut de connaître la surface exacte des organes à détecter, (Kervadec *et al.*, 2019) proposent de mettre à profit les connaissances préalables relatives à leurs limites de taille inférieure et supérieure. En fonction des applications, et donc des connaissances préalables disponibles, ces intervalles de tailles sont établis plus ou moins précisément. Quand la surface prédite n'appartient pas à l'intervalle de taille, une pénalité de type L2 est imposée à la fonction de perte. Soit $[a, b]$ l'intervalle de surface estimé à partir des connaissances préalables et $V$ la surface du masque prédit sur une image, le terme de pénalité $l_c$ s'exprime selon l'équation (1.13) :

$$l_c = \begin{cases} (V-a)^2, & \text{si } V < a \\ (V-b)^2, & \text{si } V > b \\ 0 & \text{sinon} \end{cases} \tag{1.13}$$

Les deux travaux s'assurent de la continuité et de la dérivabilité du terme $l_c$ afin que la rétropropagation soit réalisable avec l'algorithme de descente stochastique de gradient. Les pénalités mises en place dans les deux méthodes permettent d'améliorer significativement les performances et se rapprochent fortement des résultats atteints en apprentissage complètement supervisé. Pour la segmentation du ventricule gauche, l'ajout de la fonction de perte en apprentissage supervisé permet de faire passer le coefficient de Dice de 15% à 87,1%, soit seulement 2% en dessous du coefficient de Dice obtenu en apprentissage complètement supervisé.

(Kervadec *et al.*, 2019) prouvent la bonne généralisation de sa méthode en la testant sur trois applications différentes : la segmentation du ventricule gauche, de corps vertébral et de prostate sur des images par résonnance magnétique. (Kervadec *et al.*, 2019) montrent aussi que plus l'intervalle à respecter est estimé avec précision, plus les masques de segmentations prédits sont exacts.

## 1.5 Synthèse générale

La revue de littérature met en avant deux types d'approches pour réaliser la tâche de segmentation et d'identification automatique des vertèbres dans des images radiologiques.

L'utilisation directe de CNNs de segmentation d'instance pour les tâches de segmentation et d'identification de vertèbres dans des images radiologiques est possible et offre des performances solides. (Yang *et al.*, 2019) et (Kónya *et al.*, 2021) atteignent des coefficients de Dice de plus de 85% avec Mask R-CNN sur des radiographies. Les résultats des CNNs de segmentation d'instance sont significativement meilleurs que ceux des CNNs de segmentation sémantique en termes de segmentation et d'identification (Kónya *et al.*, 2021). De plus, les CNNs de segmentation d'instance sont capables de générer des résultats exacts même quand les radiographies contiennent des objets étrangers tels que des vis ou tiges d'instrumentation (Kónya *et al.*, 2021). L'inconvénient majeur de ces méthodes tient dans les phénomènes de sous- et sur-prédiction : il arrive que des vertèbres soient manquées ou que de multiples prédictions soient faites pour une même vertèbre (Wang *et al.*, 2021 ; Kónya *et al.*, 2021).

(Kónya *et al.*, 2021) proposent cependant des stratégies afin d'atténuer ce problème, notamment en utilisant la technique de NMS.

D'autres auteurs préfèrent développer leur propre structure de segmentation et d'identification de vertèbres à partir du réseau U-Net. Pour créer le cadre de segmentation d'instance nécessaire à l'obtention de bons résultats, il existe deux principales stratégies : l'application d'un processus itératif ou la division des tâches en plusieurs étapes. Le processus itératif mis en place dans la méthode (Lessmann *et al.*, 2019) permet d'identifier et de segmenter les vertèbres de manière ordonnée en suivant les vertèbres le long du rachis. (Payer *et al.*, 2020), Chen D. et (Kim *et al.*, 2021) optent pour une structure en plusieurs étapes dans laquelle les tâches de localisation, détection et segmentation sont effectuées consécutivement par différents CNNs. Les connaissances liées à la structure du rachis sont insérées dans la méthode de différentes manières, par exemple en mettant en place un processus de détection guidé (Lessmann *et al.*, 2019) ou des contraintes anatomiques à respecter Tableau 1.2.

Par rapport au contexte du projet, on peut faire la remarque générale que peu de méthodes de la littérature s'intéressent aux données radiographiques et encore moins selon la vue frontale. La plupart des méthodes utilisent des volumes CT-scans (Sekuboyina *et al.*, 2021 ; Wang *et al.*, 2021) qui, par leur nature 3D, ne présentent pas de superposition des structures et contiennent largement plus d'informations que de simples images radiographiques. Pour les méthodes qui prennent en entrée des images radiographiques, ces images sont généralement prises sous la vue latérale (Kim *et al.*, 2020 ; Kónya *et al.*, 2021). Seuls (Yang *et al.*, 2019) utilisent la vue frontale, qu'il assemble sur une même image à la vue latérale. De plus, aucune des méthodes de la littérature ne considère l'ensemble des régions anatomiques qui constituent la vertèbre : seul le corps vertébral est segmenté.

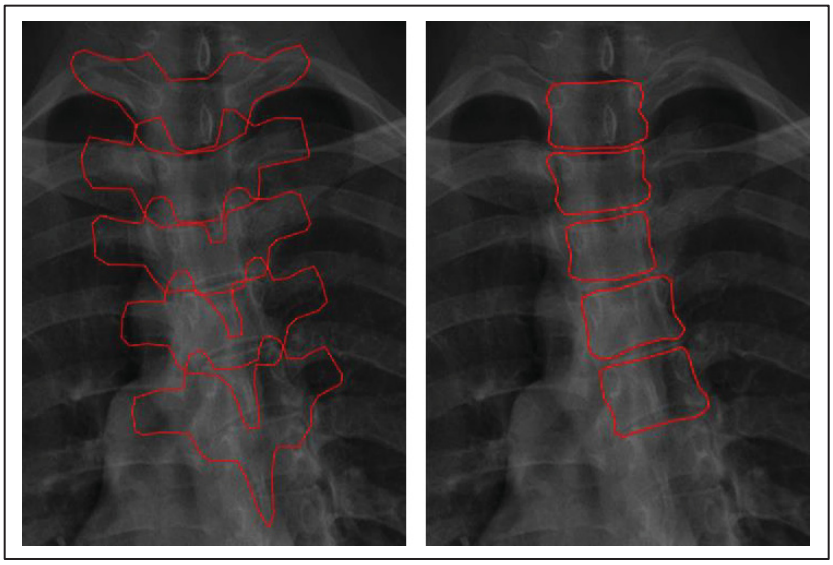


Figure 1.24 Masques de segmentation des vertèbres entières (à gauche) et des corps vertébraux (à droite) pour une même image

En considérant les vertèbres entières (soit le corps vertébral, les lames et les apophyses), ces dernières sont très superposées dans les images, notamment en vue frontale. La Figure 1.24 illustre la différence de forme des masques en considérant toute la vertèbre ou seulement le corps vertébral.

On note aussi que certaines méthodes s'intéressent seulement aux vertèbres lombaires (Kónya *et al.*, 2021 ; Kim *et al.*, 2021) et ne réalisent pas une identification individuelle de toutes les vertèbres segmentées (Kónya *et al.*, 2021 ; Yang *et al.*, 2019).

Il est particulièrement difficile de comparer les performances des différentes méthodes, car elles ne traduisent pas les mêmes résultats. En effet, à l'exception des méthodes du concours VerSe, chaque méthode utilise ses propres images d'entrée (de nature, quantité et qualité différentes), annotations (de qualité et spécifications différentes) et métriques d'évaluation. Les méthodes utilisant des CT-scans 3D sont peu comparables à des méthodes utilisant des images radiographiques 2D, les données traitées étant trop éloignées. De même, la comparaison directe des performances entre des méthodes qui segmentent toutes les vertèbres et celles qui segmentent seulement les lombaires est peu pertinente.

Bien que la comparaison directe des performances entre méthodes soit complexe, il est possible de discuter de l'intérêt, en termes de résultats, des stratégies adoptées dans les méthodes.

Par rapport aux méthodes qui utilisent simplement un CNN de segmentation d'instance, l'intérêt majeur des méthodes développées spécifiquement pour la segmentation et l'identification des vertèbres est d'intégrer de manière explicite des connaissances préalables dans le processus. La mise à profit des connaissances préalables, notamment relatives à la structure anatomique fixe et ordonnée du rachis, améliore la qualité des résultats en permettant de mieux contrôler les prédictions et en s'assurant qu'elles soient anatomiquement vraisemblables (Lessmann *et al.*, 2019 ; Payer *et al.*, 2019 ; Sekuboyina *et al.*, 2021).

Les méthodes qui consistent en une simple application d'un CNN de segmentation d'instance ne bénéficient pas de connaissances préalables explicites. Seul (Yang *et al.*, 2019) affinent ses résultats en ajoutant, en sortie du CNN, un traitement basé sur la connaissance de la forme longitudinale du rachis. L'objectif de (Yang *et al.*, 2019) étant de détecter la ligne médiane de la colonne vertébrale, seules les détections horizontalement trop éloignées sont éliminées. L'application simple d'un CNN de segmentation d'instance, bien qu'elle permette des résultats satisfaisants, n'est pas optimisée comme les méthodes développées spécialement pour la tâche de segmentation et d'identification des vertèbres le sont.

Cependant, l'utilisation d'une méthode spécifique aux vertèbres nécessite la mise en place d'une stratégie permettant la segmentation d'instance. En effet, une tâche de segmentation sémantique ne peut discerner des vertèbres superposées. La segmentation d'instance est d'autant plus nécessaire dans le cadre de ce projet, où les objets à détecter sont très superposés. Les méthodes en plusieurs étapes de (Payer *et al.*, 2021), (Kim *et al.*, 2021) et de l'équipe de Chen (Sekuboyina *et al.*, 2021) imposent l'entrainement de plusieurs réseaux et ne sont pas des solutions bout à bout. Si des erreurs se produisent lors d'une étape, les autres étapes sont complètement déstabilisées. Le processus itératif mis en place par (Lessmann *et al.*, 2019) implique la possibilité d'erreurs en cascade. De plus, (Lessmann *et al.*, 2019) préviennent que rien n'assure que le masque de segmentation couvre seulement une vertèbre si plusieurs sont visibles dans la RoI. Ils précisent aussi que ce risque est d'autant plus élevé lorsque les vertèbres sont très superposées, ce qui est le cas dans le contexte du projet. La méthode pourrait

ainsi être bien moins performante pour la détection des vertèbres entières sur des radiographies en vue frontale. Au contraire, l'utilisation d'un CNN de segmentation d'instance est directe et permet de détecter et segmenter des objets superposés avec succès (Kónya *et al.*, 2021).

À l'instar des méthodes spécialement développées pour la segmentation et l'identification (Sekuboyina *et al.*, 2021 ; Lessmann *et al.*, 2019 ; Payer *et al.*, 2021 ; Kim *et al.*, 2021), il semble essentiel d'adapter la méthode au cadre précis de la tâche à réaliser, notamment en intégrant des connaissances préalables supplémentaires. Cependant, ces méthodes, à plusieurs étapes ou itérative, sont plus contraignantes à mettre en place et à utiliser que l'application d'un CNN de segmentation d'instance. Par rapport au contexte du projet, utiliser un CNN de segmentation d'instance parait plus adapté : ils ont l'avantage d'avoir été testés sur des radiographies frontales (Yang *et al.*, 2019) et sont capables de segmenter et identifier des vertèbres superposées (Kónya *et al.*, 2021).

Dans la suite de la revue de littérature, on s'est ainsi penché sur les stratégies de spécialisation d'un CNN à un cadre d'utilisation précis. Plusieurs études proposent différents moyens d'optimiser les CNNs selon la tâche médicale désirée. L'apprentissage par transfert est plus efficace quand les images sources et cibles sont très similaires. L'étude de (Romero *et al.*, 2020) montre que l'utilisation d'images sources et cibles montrant les mêmes régions anatomiques permet d'obtenir de meilleures performances. Différents travaux proposent aussi d'intégrer aux CNNs des connaissances préalables sous forme de contraintes. Les contraintes souples, qui consistent en des termes de pénalités dans la fonction de perte, permettent aux CNNs d'apprendre des contraintes anatomiques. (Kervadec *et al.*, 2019) contraignent par exemple la taille des régions à détecter avec un intervalle à respecter. Les contraintes souples sont relativement simples à mettre en place et, bien qu'elles ne garantissent pas le respect des contraintes en tout temps, elles améliorent significativement les résultats.

# CHAPITRE 2

# PROBLÉMATIQUE ET OBJECTIFS DU PROJET

Ce projet s'inscrit dans le cadre de l'automatisation des applications de l'entreprise EOS Imaging. Plus spécifiquement, on s'intéresse à l'étape cruciale de segmentation et d'identification automatique des vertèbres sur les images radiographiques EOS. On se restreint à l'étude de la radiographie en vue frontale, qui, contrairement à la vue latérale, contient les informations nécessaires au calcul de l'angle de Cobb, métrique de référence pour le diagnostic de la scoliose. On s'intéressera aussi seulement aux vertèbres thoraciques et lombaires, qui suffisent pour reconstruire un modèle 3D de rachis utilisable en clinique.

Compte tenu du bilan de la revue de la littérature, nous choisissons d'appliquer un CNN de segmentation d'instance développé pour fonctionner sur des images naturelles à notre tâche de segmentation et d'identification individuelle des vertèbres sur des radiographies frontales EOS. En prenant aussi en considération que la spécialisation d'une méthode aux spécificités de sa tâche permet d'améliorer ses performances, nous proposons aussi de mettre en œuvre différentes techniques pour spécialiser au mieux le CNN de segmentation d'instance choisi à notre tâche. Des connaissances préalables additionnelles seront intégrées à la méthode de manière implicite à l'aide d'un pré-entrainement ajusté et de manière explicite dans un post-traitement et une fonction de perte anatomiquement contraignants.

Par rapport à la littérature, ce projet se différencie particulièrement dans le choix de l'image d'entrée et des objets à segmenter : les vertèbres entières (corps vertébral, lames et apophyses) seront identifiées et segmentées sur une radiographie frontale du rachis. La nature radiographique de l'image ainsi que la superposition accrue des objets dans la vue frontale constituent des défis encore non explorés dans la littérature. De ce fait, une étude comparative de plusieurs CNNs de segmentation d'instance populaires sera réalisée afin de choisir un CNN optimal pour cette application.

L'objectif général est de proposer une méthode de segmentation et de détection individuelle des vertèbres sur des radiographies frontales EOS® de sujets SIA.

Plus spécifiquement, le projet vise à atteindre les objectifs suivants :

- Afin de déterminer le CNN optimal pour cette application, effectuer une étude comparative des performances de différents CNNs de segmentation d'instance pour la détection et la segmentation de vertèbres dans des radiographies frontales EOS.
- À partir du CNN optimal, mettre en place et évaluer une méthode CNN spécialisée à la tâche de détection et de segmentation des vertèbres dans les radiographies frontales EOS.

# CHAPITRE 3

# ÉTUDE COMPARATIVE DE CNNS DE POINTE POUR LA DÉTECTION ET LA SEGMENTATION DES VERTÈBRES DANS DES RADIOGRAPHIES FRONTALES EOS

## 3.1 Introduction

Plusieurs études récentes appliquent directement des CNNs de segmentation d'instance populaires à la tâche de segmentation et d'identification de vertèbres sur des images radiologiques. Ces CNNs de segmentation d'instance, initialement conçus pour fonctionner sur des images naturelles, tels que Mask R-CNN et YOLACT, performent avec succès sur des données radiologiques.

Dans ce chapitre, en suivant la même stratégie, différents CNNs de segmentation d'instance sont mis en œuvre à la tâche de segmentation et d'identification individuelle de vertèbres sur des radiographies frontales EOS de patients atteints de SIA. De manière similaire à (Kónya *et al.*, 2021), une étude comparative des performances des CNNs est réalisée. En plus d'évaluer les capacités des CNNs sur notre tâche, on cherche à savoir si certains performent mieux que d'autres.

Ce chapitre commence par une présentation des données utilisées et des différents CNNs de segmentation d'instance étudiés. La méthodologie de l'étude comparative est donnée puis les résultats de cette dernière sont analysés.

## 3.2 Données et CNNs étudiés

### 3.2.1 Données

L'utilisation d'un CNN reposant considérablement sur les données d'entrée, il est essentiel de leur porter une attention particulière. En outre, cette étude présente la spécificité de faire fonctionner des CNNs sur des données éloignées de celles sur lesquelles ils sont ordinairement utilisés. Les CNNs sont généralement entrainés sur des ensembles d'images naturelles, tels que COCO ou ImageNet qui contiennent des images montrant des objets communs. Dans cette étude, les données consistent en des images radiographiques de rachis.

#### 3.2.1.1 Nature des données collectées

La base de données utilisée dans cette étude comporte 767 images radiographiques EOS du rachis en vue frontale. Ces images radiographiques ont été acquises sur des patients sains et atteints de SIA avec la cabine de radiographie biplane EOS à l'Hôpital Sainte Justine de Montréal, QC. L'accès aux données a été approuvé par les comités d'éthique de la recherche de l'Hôpital Sainte Justine de Montréal, QC; du Centre Hospitalier Universitaire de Montréal, QC et de l'École de Technologie Supérieure de Montréal, QC.

Pour chacune des acquisitions de la base de données, une reconstruction 3D du rachis a été réalisée avec le logiciel SterEOS©. Le champ visible, de même que la taille et le format varient entre les images. La hauteur des images est comprise entre 2696 et 5379 pixels tandis que la largeur varie entre 1764 et 1984 pixels. Le champ contient toujours le rachis entier et se coupe verticalement au niveau du crâne et du haut des cuisses.

Pour 747 images sur les 767 totales, des informations quant à l'angle de Cobb maximal sont disponibles. Quand une image présente un angle de Cobb maximal inférieur à 10°, sa valeur précise n'est pas renseignée. Les données avec un angle de Cobb maximal inconnu représentent 2,6% de la base de données. Sur les 747 images, la valeur maximale de l'angle de Cobb atteinte est de 123,1°.

L’ensemble de données présente 13,7% d’images postopératoires. Les images postopératoires contiennent des objets étrangers tels que des vis et des tiges métalliques d’instrumentation. On note aussi que la base de données ne contient pas de cas sur- ou sous-numéraire : toutes les images montrent 12 vertèbres thoraciques et 5 vertèbres lombaires. Des exemples d’images, avec et sans instrumentation, sont donnés en Figure 3.1.

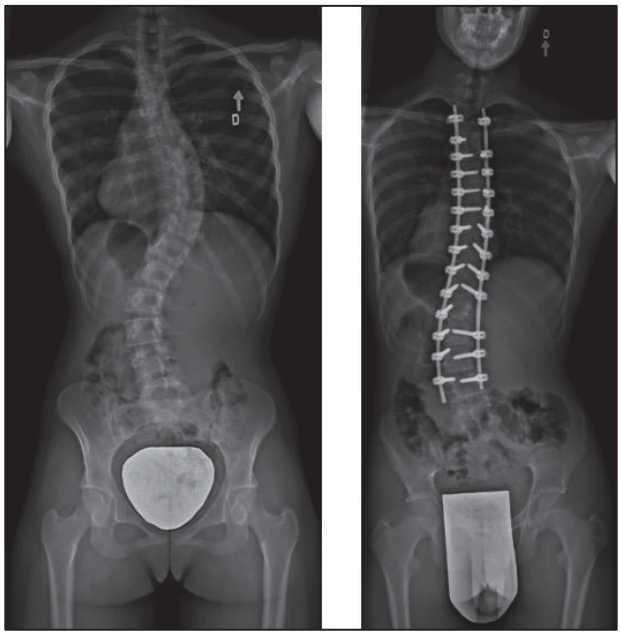


Figure 3.1 Radiographies frontales EOS de rachis atteints de SIA. À droite, l'image est postopératoire

Il est possible d’analyser la distribution de la base de données en fonction de l’angle de Cobb avec cohérence. L’histogramme de la distribution des 97,4% de la base de données dont on connait l’angle de Cobb est donné en Figure 3.2.

Pour les données dont on sait seulement qu’elles présentent un angle de Cobb inférieur à 10°, on fixe ce dernier à 10°. La moyenne de l’angle de Cobb sur les données est de 39,5°. On remarque que 95% des données présentent un angle de Cobb inférieur à 77,9°. Les derniers 5% s’étalent dans l’intervalle d’angle [78° ; 123,1°].

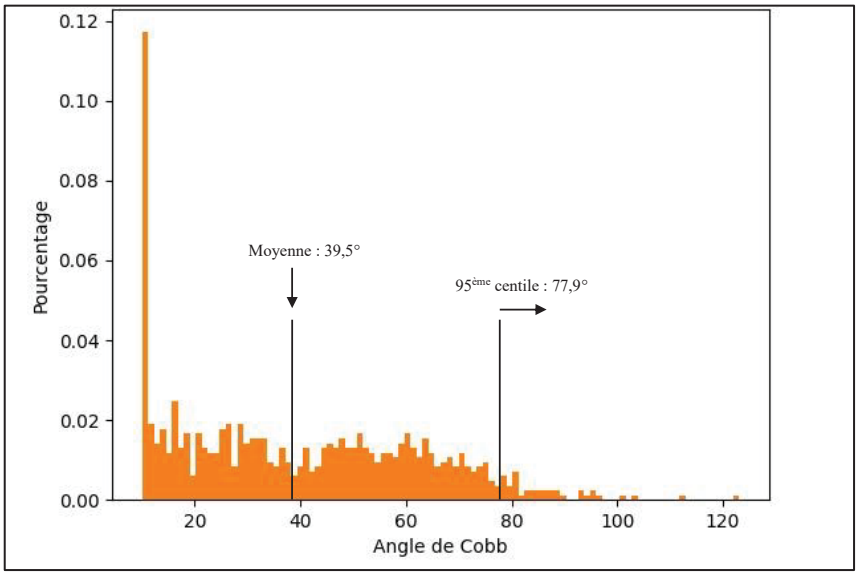



Figure 3.2 Histogramme des données en fonction de l'angle de Cobb

La SOSORT (*Scientific Society on Scoliosis Orthopaedic and Rehabilitation Treatment*) (Negrini *et al.*, 2012) met en place une classification des types de scoliose selon l'angle de Cobb pour les patients atteints d'AIS. Le Tableau 3.1 présente le pourcentage de données de chaque type de scoliose selon la classification SOSORT, caractérisant ainsi la population de la base de données.

Tableau 3.1 Classification SOSORT de l'ensemble de données. Les 2,6% de données restantes correspondent aux données dont les angles de Cobb sont inconnus

| **Type de scoliose** | **Angle de Cobb (°)** | **Données correspondantes dans la base (%)** |
|---|---|---|
| Faible | 5 - 24 | 31,3 |
| Modérée | 25 - 45 | 24,0 |
| Sévère | 45 – 59 | 19,0 |
| Très sévère | 60 et plus | 23,1 |

#### 3.2.1.2 Extraction des données de vérité de terrain

Les données de vérité de terrain, nécessaires pour réaliser l'entrainement supervisé, correspondent dans cette étude aux masques de segmentation, aux boîtes englobantes et aux labels de chaque vertèbre thoracique et lombaire sur chaque image de la base de données. Elles sont extraites depuis les reconstructions 3D préalablement réalisées pour chaque image à l'aide du logiciel commercial SterEOS©.

Les masques de segmentation de vérité de terrain correspondent aux silhouettes des modèles 3D des vertèbres dans le plan frontal. Les données de vérité terrain sont donc obtenues avec le logiciel semi-automatique SterEOS©, qui combine un processus automatique et des corrections par des opérateurs.

Les boîtes englobantes sont calculées à partir des masques de segmentation. Les labels sont, comme les masques de segmentation, directement extraits des reconstructions 3D. Les labels, au nombre de 17, correspondent aux notations médicales qui identifient les vertèbres par leur nature et leur ordre anatomique dans le rachis. Ils vont de T1 à T12 puis de L1 à L5.

Contrairement aux méthodes de segmentation en vue frontale revues qui considèrent seulement le corps vertébral, le masque de segmentation comprend dans cette étude la silhouette de la vertèbre entière : le corps vertébral ainsi que les apophyses sont discernables. La connaissance de la forme des apophyses est nécessaire pour la reconstruction 3D des rachis.

La Figure 3.3 montre la forme des masques de segmentation et des boîtes englobantes.

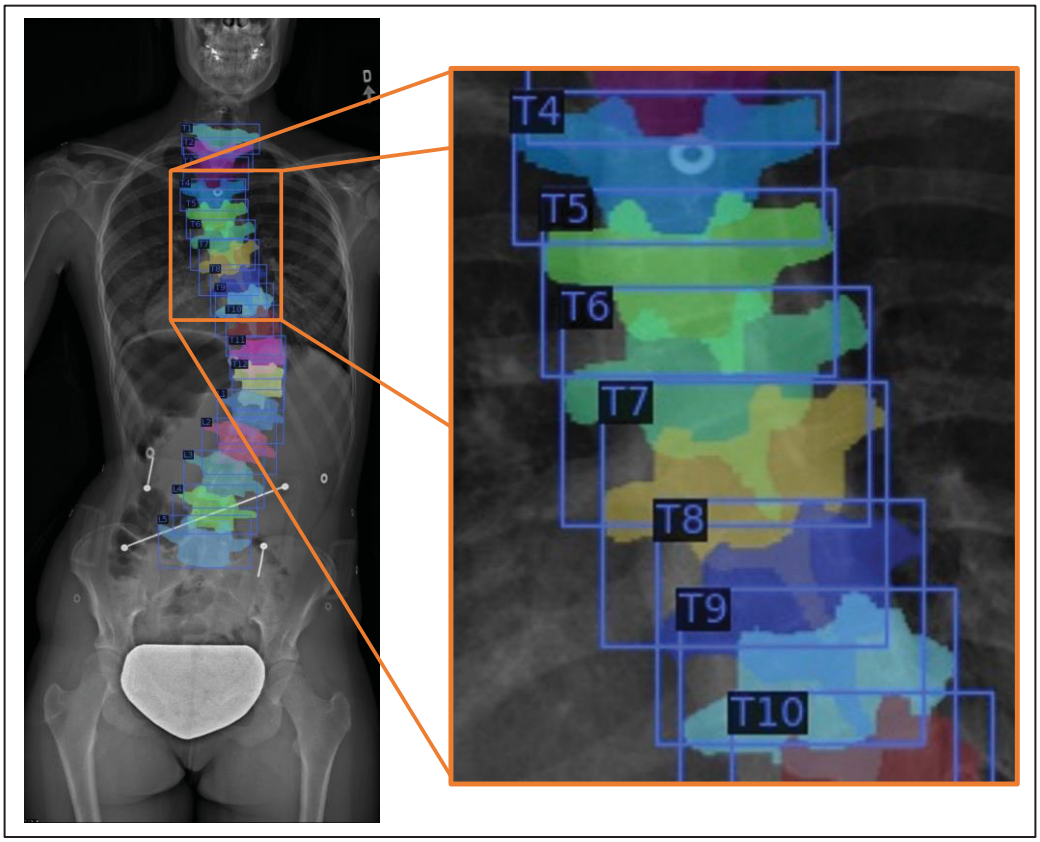



Figure 3.3 Masques de segmentation et boîtes englobantes de vérité terrain (sur l'image de gauche) sur une image radiographique avec zoom sur des vertèbres (sur l'image de droite)

### 3.2.2 CNNs de segmentation d'instance étudiés

#### 3.2.2.1 Sélection des CNNs

On décide d'étudier plusieurs modèles CNNs afin d'évaluer si les différences entre les architectures entrainent des variations de performances. Quatre modèles CNN sont sélectionnés : Mask Scoring R-CNN (MS R-CNN) (Huang *et al.*, 2019), DetectoRS (Qiao, Chen et Yuille, 2021), YOLACT (Bolya *et al.*, 2019) et RetinaMask (Fu, Shvets et Berg, 2019). Les explications relatives aux architectures de ces quatre CNNs sont données en partie 1.2.3.2. Afin d'essayer des architectures relativement différentes, deux CNNs en plusieurs étapes -MS R-CNN et DetectoRS- ainsi que deux CNNs en une étape -YOLACT et RetinaMask - sont

retenus. Le choix des CNNs est basé sur leur utilisation et performances dans la revue de littérature, mais aussi sur la disponibilité de leur code.

L'étude de (Kónya *et al.*, 2021) montre que YOLACT et Mask R-CNN sont tous deux très performants sur la tâche de segmentation et d'identification des vertèbres sur des images radiographiques latérales. Plutôt que de tester le très populaire Mask R-CNN, son amélioration MS R-CNN est sélectionnée. RetinaMask est un réseau en une étape très performant. Sur l'ensemble COCO test-dev, il surpasse YOLACT. Au début de ce projet, DetectoRS est le CNN le plus performant sur l'ensemble COCO test-dev. Les performances des quatre réseaux sur l'ensemble de test-dev sont données dans le Tableau 3.2.

Tableau 3.2 Performances des CNNs sur l'ensemble COCO test-dev. AP est la précision moyenne moyennée sur les 10 paliers d'IoU, $AP_{50}$ la précision moyenne pour un seuil d'IoU de 50% et $AP_{75}$ la précision moyenne pour un seuil d'IoU de 75%

| | **AP (%)** | **$AP_{50}$ (%)** | **$AP_{75}$ (%)** |
|---|---|---|---|
| **RetinaMask** | 34,7 | 55,4 | 36,9 |
| **MS R-CNN** | 38,3 | 58,8 | 41,5 |
| **YOLACT** | 29,8 | 48,5 | 31,2 |
| **DetectoRS** | 48,5 | 72,0 | 53,3 |

### 3.2.2.2 Implémentation des CNNs

Les codes des CNNs sélectionnés sont tous librement disponibles sur la plateforme GitHub. Les codes de YOLACT et DetectoRS sont tirés du dépôt MMDetection d'Open-MMLab (K. Chen, Wang, *et al.*, 2019). Les codes de MS R-CNN et RetinaMask sont rendus disponibles par leur auteurs sur les dépôts respectifs maskscoring_rcnn de zjhuang22 (Huang *et al.*, 2019) et retinamask de chengyangfu (Fu, Shvets et Berg, 2019). Ces deux dépôts sont basés sur le dépôt maskrcnn-benchmark de Facebook Research (Massa et Girshick, 2018).

DetectoRS et YOLACT sont construits sur le *backbone* ResNet50 (He, Xiangyu Zhang, *et al.*, 2016). MS R-CNN et RetinaMask sont respectivement construits sur les *backbones* ResNet101 (He, Xiangyu Zhang, *et al.*, 2016) et ResNeXt-101 (Xie *et al.*, 2017).

#### 3.2.2.3 Post-traitement des résultats

En sortie du CNN, on obtient des prédictions pour chaque image testée. Les prédictions consistent en des boîtes englobantes labellisées avec un masque de segmentation associé. Bien que le CNN filtre les prédictions en gardant seulement celles avec les plus hauts scores de confiance, il arrive que certaines prédictions soient redondantes. (Kónya *et al.*, 2021) le remarquaient aussi dans leur étude. Sur une image résultante, un objet à détecter (ici une vertèbre) est parfois associé à plusieurs prédictions. Par exemple, sur la Figure 3.4, le CNN propose trois prédictions labellisées « T3 » et deux prédictions labellisées « T1 ».

Comme les vertèbres sont identifiées individuellement et qu'un rachis contient toujours 17 vertèbres thoraciques et lombaires distinctes, il n'existe qu'un seul objet à détecter par catégorie. Le post-traitement consiste à éliminer les prédictions redondantes en catégorie. Pour chaque vertèbre à détecter, seulement la prédiction avec le plus haut score de confiance est gardée, comme montré sur le schéma en Figure 3.4.

Le post-traitement permet d'obtenir une image avec 17 prédictions correspondantes aux 17 vertèbres. L'évaluation est faite après application du post-traitement. Le post-traitement agit comme une opération de NMS optimisée grâce aux connaissances préalables : il y a 17 objets à détecter, chacun appartenant à une des 17 catégories.

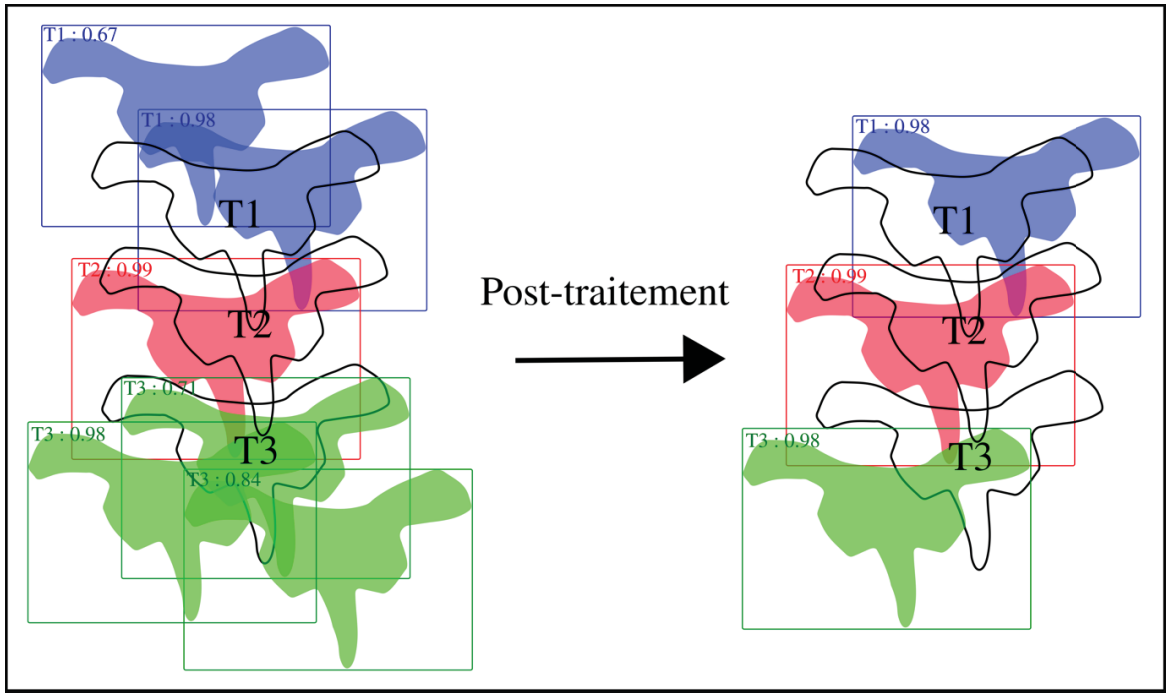



Figure 3.4 Schéma de principe du post-traitement

### 3.2.3 Outils pour l'entrainement des modèles

Les entrainements, les validations et tests des modèles sont fait sur les nœuds de calcul mis à disposition par Calcul Canada. Plus spécifiquement, toutes ces tâches sont réalisées avec un GPU NVIDIA V100 sur la grappe de calcul Graham.

## 3.3 Méthodologie

L'objectif principal de cette étude est d'évaluer les CNNs de segmentation d'instance DetectoRS, MS R-CNN, RetinaMask et YOLACT pour la tâche de segmentation et d'identification des vertèbres sur des radiographies EOS en vue frontale. La configuration optimale pour chaque CNN est déterminée grâce à une étape d'ajustement des paramètres (*fine-tuning)* réalisée en validation croisée à cinq plis. Ensuite, les quatre CNNs pris sous leur

configuration optimale sont évalués sur l'ensemble de test. Le CNN le plus performant des quatre pour la tâche est alors déterminé par analyse statistique. Enfin, une analyse poussée des résultats du CNN le plus performant est conduite.

Le travail se découpe ainsi en plusieurs étapes :

- Préparation des données;
- *Fine-tuning* des quatre CNNs sur l'ensemble d'entrainement. Le fine-tuning est réalisé selon la stratégie de validation croisée à cinq plis, comme expliqué dans la partie 1.1.5.2 de la revue de littérature. À l'issue, chaque CNN est sélectionné selon sa configuration optimale et est évalué sur l'ensemble de test;
- Comparaison des performances des quatre CNNs par analyse statistique;
- Analyse des résultats du CNN le plus performant.

### 3.3.1 Préparation des données

#### 3.3.1.1 Répartition des données dans l'ensemble de test et les plis d'entrainement

La stratégie de validation croisée à cinq plis utilisée lors de l'étape de *fine-tuning* impose de séparer l'ensemble de données en un ensemble de test et cinq sous-ensembles d'entrainement et de validation (les plis). On choisit dans cette étude de séparer l'ensemble d'entrainement et l'ensemble de test selon le rapport 85:15. L'ensemble d'entrainement et l'ensemble de test contiennent respectivement 651 et 116 images.

Les différents ensembles de données doivent être représentatifs les uns des autres pour apprendre et être évalués sur différents types de scolioses. L'angle de Cobb maximal, métrique standard pour le diagnostic de la scoliose, est un bon indice de similarité entre les images. Les images avec des angles de Cobb élevés sont visuellement très différentes des images avec des angles de Cobb faibles. Les vertèbres, qui sont les objets à identifier et segmenter, sont positionnées différemment dans l'image en fonction de l'angle de Cobb, comme illustré en Figure 3.5.

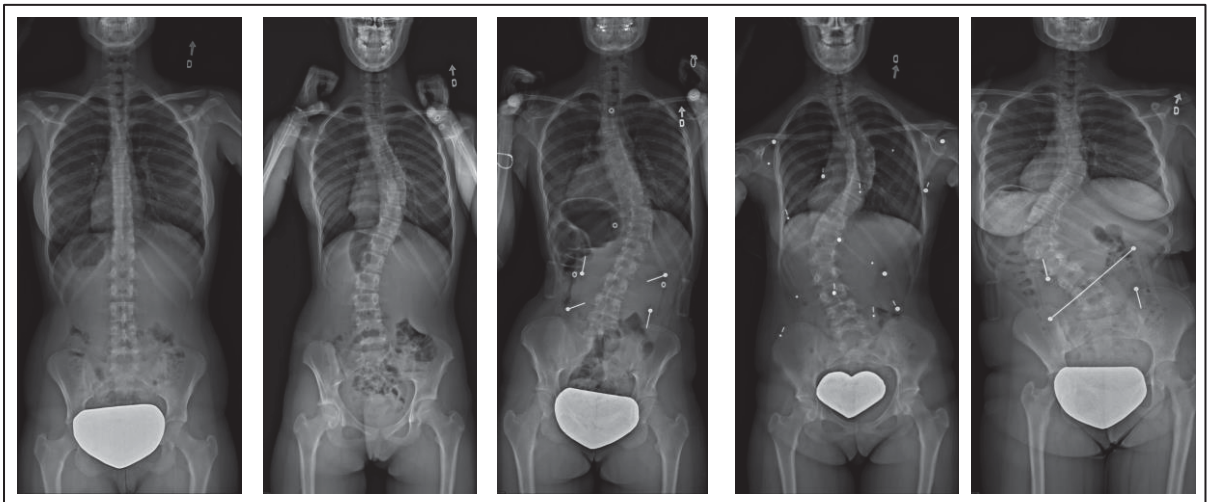



Figure 3.5 Exemples de radiographies suivant les cinq types de scoliose établis

La distribution des images dans les différents sous-ensembles est donc réalisée selon les angles de Cobb. La présence d'instrumentation ou non dans l'image n'a pas été considérée. En se basant sur la classification SORORT, donnée précédemment dans le Tableau 3.1, les données sont classées selon 4 catégories.

La distribution des cas dans la base de données incite à créer une 5$^{e}$ catégorie, la catégorie « Extrême ». Cette catégorie permet de répartir quelques données qui présentent des angles de Cobb extrêmes supérieurs à 80°. Ces cas, en plus de ne représenter que 5% de la base de données, sont visuellement très différents des autres images. Il est donc nécessaire qu'ils soient également répartis dans les ensembles.

Enfin, une 6$^{e}$ catégorie, appelée « Inconnu », rassemble les données dont l'angle de Cobb n'est pas renseigné. La distribution des images selon leur catégorie de scoliose est donnée en Figure 3.6. La distribution des images dans les différents sous-ensembles est réalisée de manière à suivre la distribution des catégories dans l'ensemble de données.

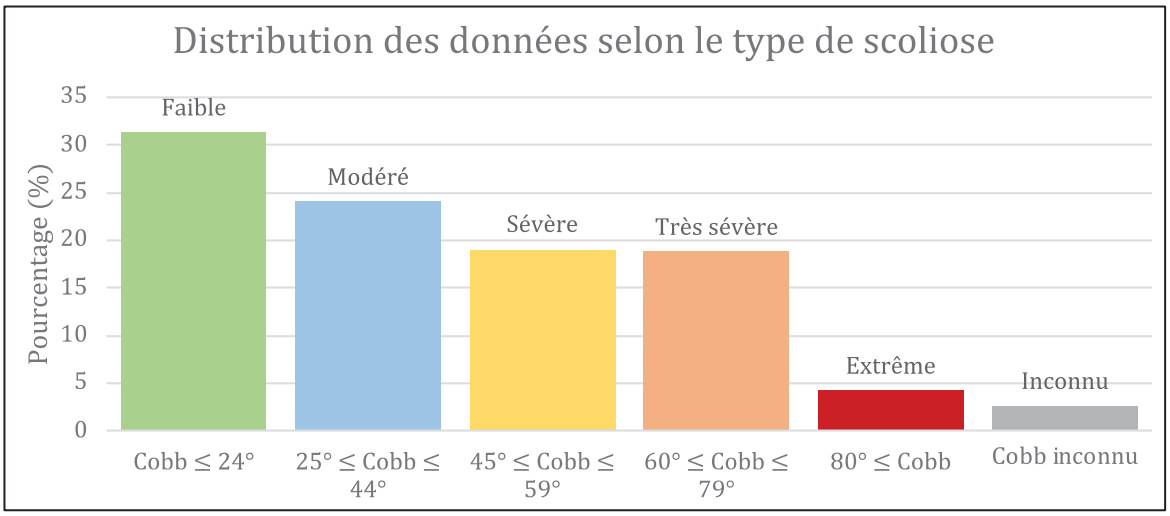



Figure 3.6 Distribution des images radiographiques selon les 6 catégories

### 3.3.1.2 Uniformisation des prétraitements des images en entrée des CNNs

Avant d'entrainer les réseaux sur les données, il est nécessaire de s'assurer que toutes les images en entrée des CNNs subissent les mêmes prétraitements.

Les prétraitements consistent en :

- Le redimensionnement des images;
- La normalisation des images;
- La transformation par renversement horizontal (*horizontal flip)* des images.

Les réseaux étant pré-entrainés sur ImageNet, la taille d'image d'entrée standard est 1333×1333 pixels. Pour profiter au maximum de l'apprentissage par transfert, les images sont redimensionnées selon la taille standard tout en préservant leur rapport de forme : leur hauteur est fixée à 1333 pixels et le nombre de pixels en largeur est calculé pour conserver le rapport de forme.

La normalisation des images correspond à la normalisation des pixels sur les trois canaux, comme réalisé sur le dépôt MMDetection (K. Chen, Wang, *et al.*, 2019). Soit une valeur de canal $x$ d'un pixel d'une image, $\tilde{X}$ la valeur moyenne du canal calculée sur tous les pixels de toutes les images de l'ensemble d'entrainement et $X_\sigma$ l'écart-type.

Le calcul de la valeur du canal normalisée $x'$se fait selon la formule (3.1):

$$x' = \frac{(x - \tilde{X})}{X_\sigma} \tag{3.1}$$

Sur les dépôts, les valeurs standard de $\tilde{X}$ et $X_\sigma$ sont calculées sur ImageNet. Les images radiographiques de notre ensemble de données étant en niveaux de gris, les valeurs sur les trois canaux d'un pixel sont égales et sont largement différentes des valeurs $\tilde{X}$ et $X_\sigma$ d'ImageNet. De ce fait, les valeurs standard $\tilde{X}$ et $X_\sigma$ sont remplacées par celles calculées sur l'ensemble d'entrainement : $\tilde{X} = 59{,}22$ et $X_\sigma = 11{,}47$ . Elles sont répétées sur les trois canaux.

La transformation par renversement horizontal est effectuée avec un taux de probabilité de 50%. Cette transformation ne crée pas d'image, mais les modifie. Elle permet de donner aux CNNs davantage d'images présentant des scolioses avec une courbure vers la gauche, qui sont largement moins fréquentes que les scolioses avec une courbure vers la droite (Choudhry, Ahmad et Verma, 2016). Cette surreprésentation des cas de scoliose avec une courbure vers la gauche les rend plus faciles à apprendre. La Figure 3.7 illustre le principe de transformation par renversement horizontal.

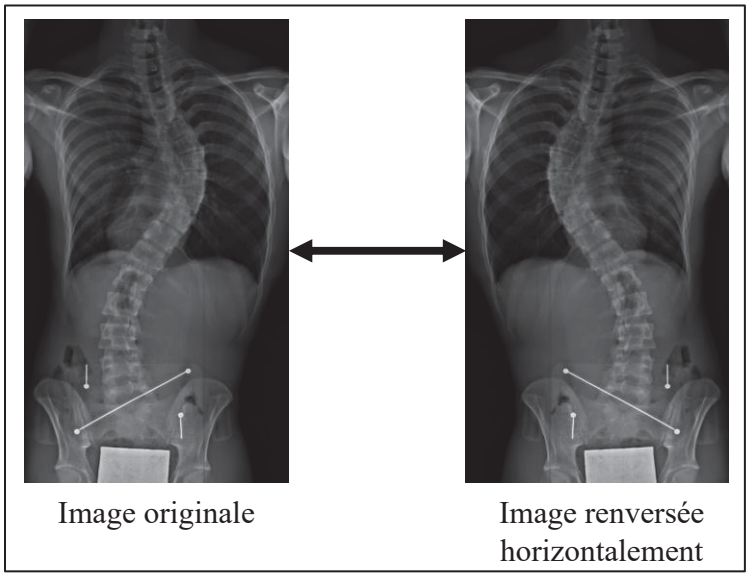



Figure 3.7 Prétraitement renversement horizontal

### 3.3.2 Métriques d'évaluation des CNNs

Sur les dépôts GitHub, les métriques d'évaluation implémentées sont celles communément utilisées pour la détection d'instance sur des images naturelles présentes dans COCO. Les métriques AP (*Average Precision*), $AP_{50}$ et $AP_{75}$ évaluent la qualité des boîtes englobantes et des masques de segmentation. Dans le cadre de cette étude, ces métriques ne sont pas particulièrement pertinentes. Bien qu'elles reflètent les performances des CNNs, elles ne correspondent pas aux métriques utilisées pour évaluer la tâche de segmentation et d'identification sur des images médicales. Afin de pouvoir comparer les résultats de cette étude à la littérature, mais aussi pour pouvoir mieux analyser les performances en segmentation et en identification, une méthode d'évaluation médicale est mise en place.

Dans le domaine médical, les tâches de segmentation et d'identification sont respectivement évaluées avec le coefficient de Dice et le taux d'identification (Rizwan I Haque et Neubert, 2020).

Le taux d'identification implique une condition d'identification : si cette condition est respectée, l'identification est réussie et inversement.

La métrique du coefficient de Dice possède une définition précise et stricte. Au contraire, le taux d'identification ne repose pas sur une définition spécifique : chaque article de la littérature propose sa propre caractérisation du taux d'identification.

Pour (Kim *et al.*, 2021) et (Kónya *et al.*, 2021), le taux d'identification qualifie l'identification d'une image entière. La condition d'identification est basée sur la présence des prédictions de chaque vertèbre dans l'image. Pour (Yang *et al.*, 2019) et le concours VerSe (Sekuboyina *et al.*, 2021), le taux d'identification caractérise l'identification des vertèbres considérées individuellement. La condition d'identification repose sur la précision de localisation de la vertèbre prédite.

Dans le contexte de cette étude, le taux d'identification renseignant sur l'identification individuelle des vertèbres est plus pertinent qu'un taux d'identification renseignant simplement sur l'existence de prédictions pour chaque vertèbre. La présence d'une vertèbre prédite dans une image n'est pas suffisante pour assurer sa bonne identification : il est nécessaire que la

vertèbre prédite soit correctement localisée. Le taux d'identification renseignant sur l'identification individuelle des vertèbres est une métrique plus complète.

La condition d'identification proposée par (Yang *et al.*, 2019) consiste en une valeur minimale de coefficient de Dice de 50% à atteindre. Dans le concours VerSe (Sekuboyina *et al.*, 2021), deux conditions sont à satisfaire : en calculant les distances entre les vertèbres par rapport à leur centroïde, une vertèbre prédite est considérée comme bien identifiée si la vertèbre de vérité terrain la plus proche est celle de catégorie correspondante et si la distance entre ces deux vertèbres est inférieure à 20 mm. Les conditions de localisation basées sur le score de Dice et sur les distances entre les centroïdes semblent toutes les deux pertinentes.

La métrique du coefficient de Dice est utilisée dans l'étude pour évaluer la tâche de segmentation. En se basant sur les taux d'identification de (Yang *et al.*, 2019) et VerSe (Sekuboyina *et al.*, 2021), quatre taux d'identification IR (*Identification Rate*) selon différentes conditions d'identification sont proposées :

- $IR_{D\geq 50}$ : une vertèbre prédite est bien identifiée quand son coefficient de Dice est supérieur ou égal à 50%;
- $IR_{D\geq 75}$ : une vertèbre prédite est bien identifiée quand son coefficient de Dice est supérieur ou égal à 75%;
- $IR_{d\leq 10}$ : en calculant les distances entre les vertèbres par rapport à leur centroïde, une vertèbre prédite est considérée comme bien identifiée si la vertèbre de vérité terrain la plus proche est celle de catégorie correspondante et si une distance inférieure ou égale à 10 mm sépare ces deux vertèbres;
- $IR_{d\leq 5}$ : en calculant les distances entre les vertèbres par rapport à leur centroïde, une vertèbre prédite est considérée comme bien identifiée si la vertèbre de vérité terrain la plus proche est celle de catégorie correspondante et si une distance inférieure ou égale à 5 mm sépare ces deux vertèbres.

Les deux types de conditions de localisation (selon Dice ou la distance) sont représentés afin d'évaluer s'ils mènent à des résultats similaires ou éloignés. Calculer $IR_{D\geq 50}$ permet de pouvoir comparer les résultats d'identification de cette étude à ceux de (Yang *et al.*, 2019), qui utilise

aussi des images EOS. La superposition entre les vertèbres voisines étant relativement élevée sur les images radiographiques de la base de données, le taux de superposition est calculé. Comme illustré sur la Figure 3.8, il représente le rapport entre la surface de recouvrement d'une vertèbre A par une vertèbre adjacente B et la surface totale de la vertèbre A.

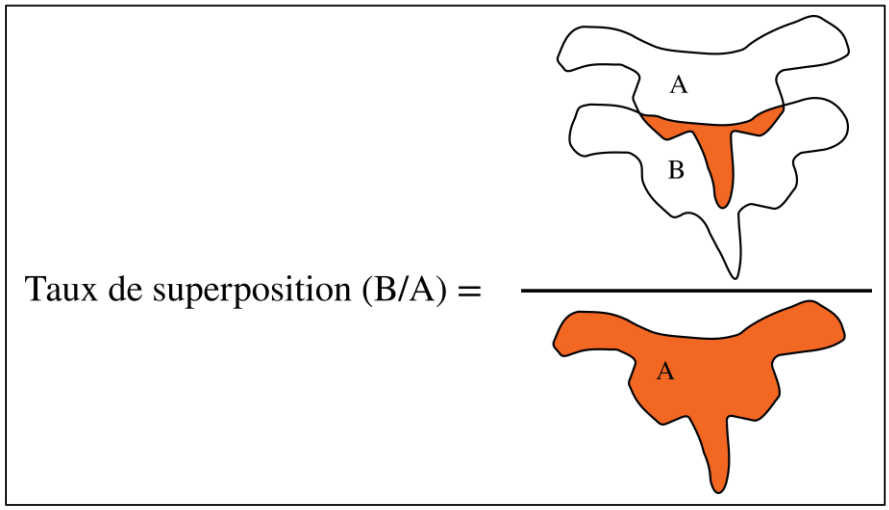



Figure 3.8 Schéma de principe du calcul du taux de superposition

Le taux de superposition est en moyenne de 15,6% et atteint une valeur maximale de 72,1%. Dans le cas où le taux de superposition de vertèbres adjacentes dépasse 50%, la condition d'identification de coefficient de Dice supérieur ou égal à 50% n'est pas suffisante. De ce fait, la métrique $IR_{D\geq 75}$, requérant un coefficient de Dice supérieur ou égal à 75%, est aussi mise en place.

La condition de distance entre les centroïdes inférieure à 20 mm utilisée par VerSe (Sekuboyina *et al.*, 2021) est aussi trop permissive dans le cadre de ce travail. Sur la base de données, la distance entre les vertèbres est calculée à partir de leurs centroïdes et en tenant compte de la résolution des images. À partir de la base de données, on estime à 24,3 mm la distance euclidienne moyenne entre deux vertèbres consécutives. Entre les vertèbres T1 et T2, la distance moyenne est de seulement 17,2 mm et atteint même une valeur minimale de 3,44 mm dans un cas extrême. Le Tableau 3.3 rassemble des informations caractéristiques des distances intervertébrales sur la base de données pour les vertèbres thoraciques et lombaires.

Tableau 3.3 Caractérisation des distances intervertébrales pour les vertèbres thoraciques et lombaires

| Distance intervertébrale (mm) | Moyenne ± Écart-type | [Min; Max] |
|---|---|---|
| Thoraciques (T1 – T12) | 21,49 ± 2,01 | [3,44 ; 36,27] |
| Lombaires (L1 – L5) | 28,11± 0,36 | [8,37; 38,60] |

Les métriques $IR_{d\leq10}$ et $IR_{d\leq5}$, qui exigent une distance entre les centroïdes maximale de respectivement 10 mm et 5 mm, sont utilisées. L'intérêt d'utiliser des taux d'identification avec des conditions plus ou moins difficiles à atteindre est de pouvoir analyser l'écart de performances qu'elles présentent.

### 3.3.3 *Fine-tuning* des réseaux

Bien qu'il soit possible de directement appliquer les CNNs sous leur configuration standard, une opération de *fine-tuning* est effectuée : on ajuste les hyperparamètres des CNNs pré-entrainés sur ImageNet. Pour chaque CNN, plusieurs modèles correspondant à différentes configurations d'hyperparamètres sont essayés. Cet ajustement permet de trouver la configuration optimale qui offre les meilleurs résultats sur les données.

#### 3.3.3.1 Stratégie de validation croisée à cinq plis

Plutôt que simplement effectuer le *fine-tuning* sur des ensembles d'entrainement et de validation uniques, on préfère utiliser la stratégie de répartition des données en validation croisée en *k* plis, expliquée dans la partie 1.1.5.2. et illustrée dans la Figure 1.7.
Compte tenu de la taille de l'ensemble de données et des coûts de calcul, on choisit dans cette étude une valeur $k = 5$. Chaque configuration de chaque modèle CNN est entrainée et validée selon la stratégie de validation croisée en 5 plis. La configuration optimale de chaque CNN est

alors déterminée. Chaque CNN, pris sous sa configuration optimale, est alors entrainé sur l'ensemble d'entrainement entier, soit les 5 plis, et est testé sur l'ensemble de test.

#### 3.3.3.2 Hyperparamètres ajustés

Les réseaux très profonds présentant un nombre important d'hyperparamètres, il est impensable de tous les ajuster. De plus, les dépôts présentent les CNNs selon les configurations précisées par leurs auteurs : les CNNs sont déjà optimisés pour la segmentation d'instance sur des images naturelles.

Dans ce travail, l'ajustement des hyperparamètres se limitera à l'optimisation du taux d'apprentissage $\eta$ (*learning rate*) et de la taille de lot $B$ (*batch size*) considéré à chaque itération. Le taux d'apprentissage $\eta$ est considéré comme un des hyperparamètres les plus importants, son ajustement doit toujours être réalisé (Bengio, 2012). Un taux d'apprentissage ajusté correspond au plus grand taux d'apprentissage qui n'entraine pas la divergence du CNN (Bengio, 2012). La taille de lot correspond au nombre d'échantillons utilisés pour estimer le gradient d'erreur. C'est un hyperparamètre important qui régule l'exactitude de l'estimation de l'erreur lors de l'entrainement du réseau (Brownlee, 2019c). Une taille de lot restreinte provoque généralement une mise à jour de poids bruitée permettant un apprentissage rapide et robuste (Brownlee, 2019c).

Les résultats en validation croisée à cinq plis obtenus sur une configuration guident la détermination des valeurs hyperparamètres $\eta$ et $B$ constituant une nouvelle configuration. Le *fine-tuning* est réalisé en partant des configurations standard données sur les dépôts propres à chaque CNN. Il est tout de même nécessaire de modifier la taille de lot : les images étant relativement grandes, les calculs ne peuvent s'effectuer qu'avec une taille de lot maximale de $B = 2$. L'hyperparamètre de taille de lot testé $B$ varie donc entre 1 et 2. Le taux d'apprentissage testé $\eta$ varie entre 0,001 et 0,01. La configuration de base est : $B = 2$ et $\eta = 0{,}0025$. Les entrainements des CNNs sont arrêtés quand le point de stabilité est atteint afin d'éviter le phénomène de surapprentissage.

#### 3.3.3.3 Sélection de la configuration optimale

Le *fine-tuning* des CNNs sous plusieurs configurations d'hyperparamètres permet de connaître leurs performances sur les cinq métriques d'évaluation : Dice, $IR_{D\geq 50}$, $IR_{D\geq 75}$, $IR_{d\leq 10}$ et $IR_{d\leq 5}$. Afin de sélectionner la meilleure configuration pour chaque CNN, il est nécessaire de mettre en place un score traduisant les performances globales d'un CNN sur les cinq métriques. Ce score prend en compte les valeurs moyennes des cinq métriques selon l'équation (3.2).

$$Score = \frac{4 * Dice + IR_{D\geq 50} + IR_{D\geq 75} + IR_{\leq 10} + IR_{\leq 5}}{8} \quad (3.2)$$

Le coefficient de Dice est pondéré par 4 afin que l'évaluation de la tâche de segmentation soit autant considérée que l'évaluation de la tâche d'identification. Le score est calculé pour chaque configuration d'un CNN.

### 3.3.4 Comparaison des CNNs optimaux par analyses statistiques

Chaque modèle CNN pris sous sa configuration optimale est évalué sur l'ensemble de test. Cette évaluation sur un ensemble de données qui leur est totalement inconnu révèle les performances réelles des CNNs. L'objectif est de comparer les performances des CNNs en observant s'ils présentent des différences significatives sur les métriques d'évaluation.

Le test de Shapiro-Wilk (Shapiro et Wilk, 1965) est utilisé pour vérifier si les résultats des images tests suivent une distribution normale sur chaque métrique et chaque CNN. Le test non paramétrique de Kruskal-Wallis (Kruskal et Wallis, 1952) pour échantillons indépendants est ensuite conduit. Le test est réalisé sur les résultats des quatre CNNs pour chaque métrique. Il teste l'hypothèse $H_0$ que les échantillons proviennent d'une même distribution. Lorsque, sur une métrique, des différences significatives existent entre les CNNs les intervalles de Bonferroni (Bonferroni, 1936) à 95% post-hoc sont calculés par paire. Ils indiquent si des différences significatives existent entre deux CNNs.

### 3.3.5 Analyse complémentaire des résultats du CNN le plus performant

Une analyse poussée des résultats du CNN le plus performant est nécessaire afin de mieux connaître ses capacités et son comportement. Cette analyse est conduite sur les résultats en validation croisée à cinq plis et non pas sur l'ensemble de test afin de ne pas créer de fuite des données (*data leakage*). Cette analyse comprend plusieurs études : l'étude de l'influence des données d'entrée, l'étude de la distribution des résultats, l'étude visuelle des erreurs récurrentes et l'étude de la sélection des prédictions.

#### 3.3.5.1 Analyse de l'influence des données d'entrée

L'influence de trois facteurs liés aux données d'entrée est étudiée sur les résultats du CNN : le type de vertèbre, le type de scoliose et le type d'image (postopératoire ou non).
On cherche à savoir si certaines vertèbres du rachis sont mieux identifiées et segmentées que d'autres en calculant les performances par rapport à chacune des 17 vertèbres du rachis.

#### 3.3.5.2 Analyse de la distribution des résultats

Ensuite, afin de mieux apprécier les résultats du CNN, il est nécessaire d'analyser la distribution des résultats du CNN sur les métriques d'évaluation. La métrique de taux de succès (SR pour *Success Rate*) est introduite. Elle se calcule sur chaque image : si toutes les vertèbres sont bien identifiées dans l'image, l'image est en succès, sinon, l'image est en échec. Soit $SR_C$ le taux de succès selon la condition d'identification $C$ sur un ensemble de données de $N$ images, $I_{C,i}$ l'identification selon la condition $C$ d'une vertèbre $i$ sur une image. $SR_C$ est calculé selon l'équation (3.3).

$$SR_C = \frac{1}{N}\sum_{n=1}^{N} S_n \quad \text{où } S_n = \begin{cases} 1 & si\ I_{C,i} = 1\ \ \forall\, i \in [T1, L5] \\ 0 & sinon \end{cases} \tag{3.3}$$

La métrique $SR_C$ indique donc la proportion d'images complètement bien identifiées selon une condition $C$. Elle présente un réel intérêt dans le contexte médical, car elle signale si les résultats d'une image sont utilisables ou non. La métrique de taux de succès $SR_C$ est calculée sous quatre versions, correspondant à chacune des métriques d'identification : $SR_{D\geq 50}$, $SR_{D\geq 75}$, $SR_{d\leq 10}$ et $SR_{d\leq 5}$.

#### 3.3.5.3 Analyse visuelle des images résultantes

L'analyse visuelle des images résultantes est conduite pour déterminer les erreurs récurrentes faites par le CNN. On cherche à qualifier leur nature, leur apparition, leurs causes et conséquences sur les résultats.

#### 3.3.5.4 Analyse de la sélection des prédictions

Enfin, la capacité du CNN à mettre en avant les meilleures prédictions est étudiée. En sortie du CNN, les images résultantes sont traitées par un post-traitement. Ce dernier permet de garder seulement une prédiction par catégorie de vertèbre, celle dont le score de confiance est le plus haut. Le score de confiance, qui est en fait la probabilité de classification estimée par le CNN, est l'unique indice permettant de comparer les prédictions d'une même catégorie entre-elles. On cherche à savoir si ce score de confiance reflète réellement la qualité d'une prédiction. On définit la meilleure prédiction pour une vertèbre de vérité terrain comme la prédiction de catégorie correspondante ayant le coefficient de Dice le plus élevé parmi toutes celles proposées par le CNN. Est-ce que les prédictions de plus haut score de confiance correspondent véritablement aux meilleures prédictions disponibles ?

Afin de répondre à cette question, les résultats du CNN avec post-traitement sont comparés à ceux du CNN où les meilleures prédictions sont sélectionnées avec connaissance des données de vérité terrain.

## 3.4 Résultats et comparaison des quatre CNNs

### 3.4.1 *Fine-tuning* des réseaux

#### 3.4.1.1 Entrainements sur les différentes configurations

Les courbes de perte en entrainement et en validation sont calculées pour les CNNs DetectoRS, MS R-CNN et YOLACT sur les cinq plis des différentes configurations testées. Le code de RetinaMask ne permettant pas de calculer la perte en validation au cours de l'entrainement, seule la courbe de perte en entrainement est disponible.

Des exemples de courbes de perte en entrainement et en validation calculés sur un plis d'une configuration pour chaque CNN sont donnés ci-dessous sur la Figure 3.9. À cause des différentes échelles utilisées, ces courbes ne sont pas comparables entre elles. Il est cependant possible d'apprécier individuellement leur allure.

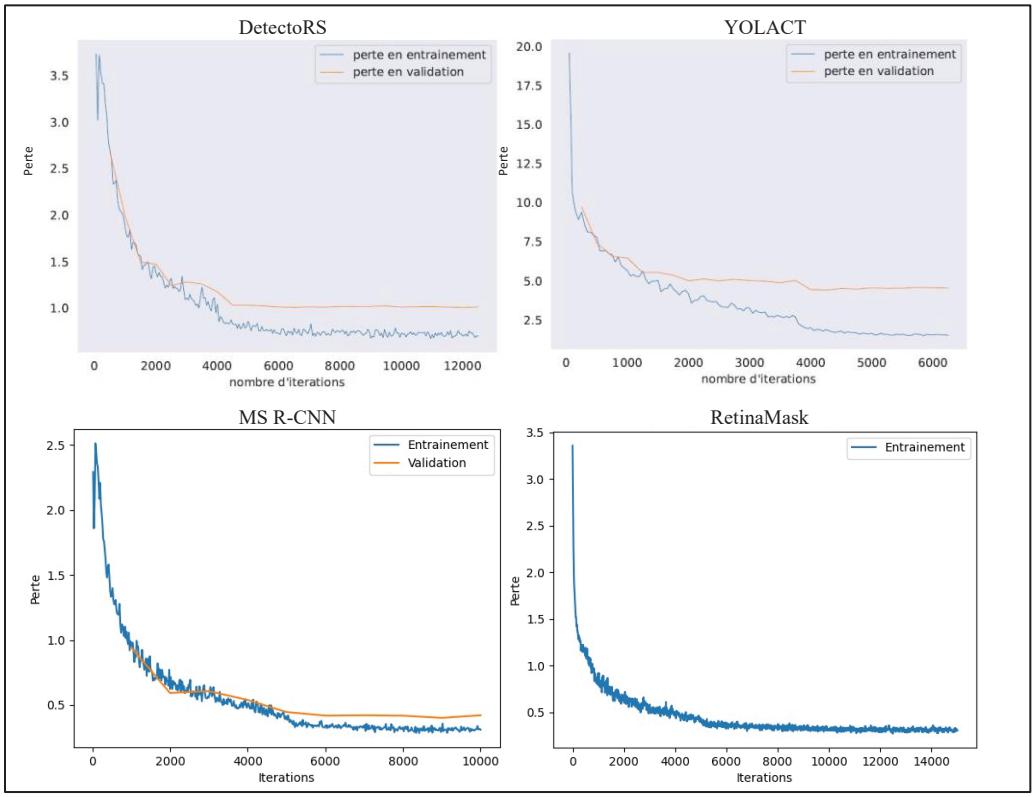



Figure 3.9 Graphiques de perte en entrainement et en validation pour les quatre CNNs

Les quatre courbes de perte en entrainement diminuent toutes au cours des itérations jusqu'à atteindre un palier. Ce palier se situe pour les quatre modèles aux alentours de 6000 itérations. Pour DetectoRS, YOLACT et MS R-CNN, on voit que la courbe de perte en validation présente la même allure que la courbe de perte en entrainement tout en maintenant un décalage vertical. YOLACT présente un écart assez prononcé entre les courbes de perte en entrainement et validation.

Le *fine-tuning* des réseaux résulte en l'entrainement et la validation de quatre configurations différentes de DetectoRS et de cinq configurations différentes pour chacun des trois autres CNNs. Il arrive parfois que, sur certaines configurations, l'entrainement d'un CNN sur un plis diverge et s'arrête. C'est le cas pour DetectoRS lorsqu'un taux d'apprentissage $\eta$=0,0075 et une taille de lot B=1 sont utilisés. De même MS-RCNN diverge pour les configurations $\eta$=0,0050; B=1 et $\eta$=0,0075; B=2.

Entre la configuration standard ($\eta$ = 0,0025 ; B = 2) et la configuration optimisée, les gains de performances moyennés sur les quatre CNNs sont de 0,7 pp pour Dice, de 0,37 pp pour $IR_{D\geq50}$, de 1,09 pp pour $IR_{D\geq75}$, de 0,6 pp pour $IR_{d\leq10}$, et de 1,69 pp pour $IR_{d\leq5}$. L'ensemble de ces configurations testées pour chaque CNN ainsi que leurs résultats moyens de validation croisée à 5 plis sur les métriques d'évaluation sont indiqués en ANNEXE I.

Les configurations optimisées des quatre CNNs sont déterminées avec le calcul du score selon l'équation (3.2). Ces configurations ainsi que leurs performances associées sont données dans le Tableau 3.4.

Tableau 3.4 Résultats des quatre CNNs en validation selon leur configuration optimisée

| | Métriques d'évaluation (moyenne ± écart-type sur les 5 plis) (%) | | | | | Score |
|---|---|---|---|---|---|---|
| CNN - Configuration | Dice | $IR_{D\geq50}$ | $IR_{D\geq75}$ | $IR_{d\leq10}$ | $IR_{d\leq5}$ | |
| DetectoRS: $\eta$ = 0,0050 ; B = 1 | 85,89±0,94 | 95,58±1,45 | 94,00±1,52 | 95,21±1,28 | 90,81±1,17 | 89,89 |
| MS R-CNN: $\eta$ = 0,0050 ; B = 2 | 85,13±0,86 | 95,37±1,43 | 92,07±1,46 | 94,07±1,43 | 85,16±2,04 | 88,40 |
| RetinaMask: $\eta$ = 0,0075 ; B = 1 | 85,57±1,08 | 94,92±1,63 | 92,81±1,70 | 94,32±1,67 | 87,68±2,15 | 89,00 |
| YOLACT: $\eta$ = 0,0075 ; B = 2 | 79,94±0,78 | 90,28±1,19 | 80,28±1,44 | 83,17±1,16 | 65,13±1,60 | 79,83 |

Au regard des résultats du Tableau 3.4, DetectoRS est le réseau le plus performant avec un score total de 89,89. Il est suivi par RetinaMask et MS R-CNN qui affichent respectivement des scores de 89 et 88,4. YOLACT, présentant un score de 79,83, performe nettement moins bien que les trois autres réseaux. L'écart de performance entre DetectoRS et les autres CNNs est plus marqué sur $IR_{D\geq 75}$ et $IR_{d\leq 5}$.

Ces résultats, obtenus en validation croisée à cinq plis, sont intermédiaires. Il est encore nécessaire de tester les CNNs selon les configurations déterminées sur un ensemble de données test, inconnu des CNNs.

### 3.4.1.2 Résultats des CNNs optimisés sur l'ensemble de test

Les résultats des quatre CNNS entrainés sur l'ensemble d'entrainement entier et évalués sur l'ensemble de test sont donnés dans le Tableau 3.5. Le score est aussi calculé suivant l'équation (3.2).

Tableau 3.5 Résultats des quatre CNNs sur l'ensemble de test

| | Métriques d'évaluation (%) (moyenne) | | | | | Score |
|---|---|---|---|---|---|---|
| CNN | Dice | $IR_{D\geq 50}$ | $IR_{D\geq 75}$ | $IR_{d\leq 10}$ | $IR_{d\leq 5}$ | |
| DetectoRS | 86,36 | 95,79 | 94,88 | 95,54 | 92,09 | 90,47 |
| MS R-CNN | 84,22 | 94,17 | 90,72 | 93,1 | 83,62 | 87,31 |
| RetinaMask | 84,59 | 93,36 | 91,73 | 92,8 | 86,61 | 87,86 |
| YOLACT | 79,49 | 88,44 | 80,17 | 82,15 | 64,3 | 79,13 |

Sur l'ensemble de test, DetectoRS obtient le score le plus élevé. Il dépasse le deuxième meilleur CNN, RetinaMask, de 2,61 points. MS R-CNN et RetinaMask montrent des scores similaires alors que YOLACT est en retrait. Une brève analyse visuelle des images résultantes de YOLACT montre que le réseau produit des masques de segmentation dont les formes sont très approximatives.

DetectoRS est aussi le plus performant sur chacune des métriques d'évaluation. Il offre un coefficient de Dice de 86,36% et des taux d'identification compris entre 92,09% pour $IR_{d\leq5}$ et 95,79% pour $IR_{D\geq50}$. Il dépasse RetinaMask de 3,15 pp sur $IR_{D\geq75}$ et de 5,48 pp sur $IR_{d\leq5}$.

### 3.4.2 Comparaison des CNNs optimaux par analyses statistiques

Pour toutes les métriques des quatre CNNs, l'hypothèse de normalité des résultats peut être rejetée avec un niveau de confiance de 95% selon le test de Shapiro-Wilk.
Avec le test non paramétrique de Kruskal-Wallis, des différences significatives de distribution sont trouvées entre les résultats des 4 CNNs sur chacune des métriques d'évaluation. Les intervalles de Bonferroni calculés montrent que YOLACT est significativement moins performant que MS R-CNN, RetinaMask et DetectoRS sur toutes les métriques. Deux autres différences significatives sont remarquées avec les intervalles de Bonferroni : DetectoRS est significativement meilleur que RetinaMask et MS-RCNN sur la métrique $IR_{d\leq5}$. L'ensemble des résultats des tests statistiques est donné en ANNEXE II.

## 3.5 Résultats complémentaires de DetectoRS

Les précédents résultats montrent que DetectoRS est particulièrement performant pour les tâches de segmentation et d'identification de vertèbres sur des radiographies EOS de rachis en vue frontale. Afin de mieux apprécier les capacités de DetectoRS, des résultats complémentaires sont tirés.

### 3.5.1 Résultats selon les données

À partir des résultats de validation sur DetectoRS, une analyse des performances par rapport aux différentes vertèbres, à l'angle de Cobb et au type d'image est conduite. On cherche à connaître l'influence sur les résultats de la gravité de la scoliose, de la présence d'outils d'instrumentation et de la catégorie de la vertèbre détectée.

#### 3.5.1.1 Résultats selon les vertèbres

Les résultats de validation sur les cinq métriques sont calculés en fonction des vertèbres et affichés en Figure 3.10.

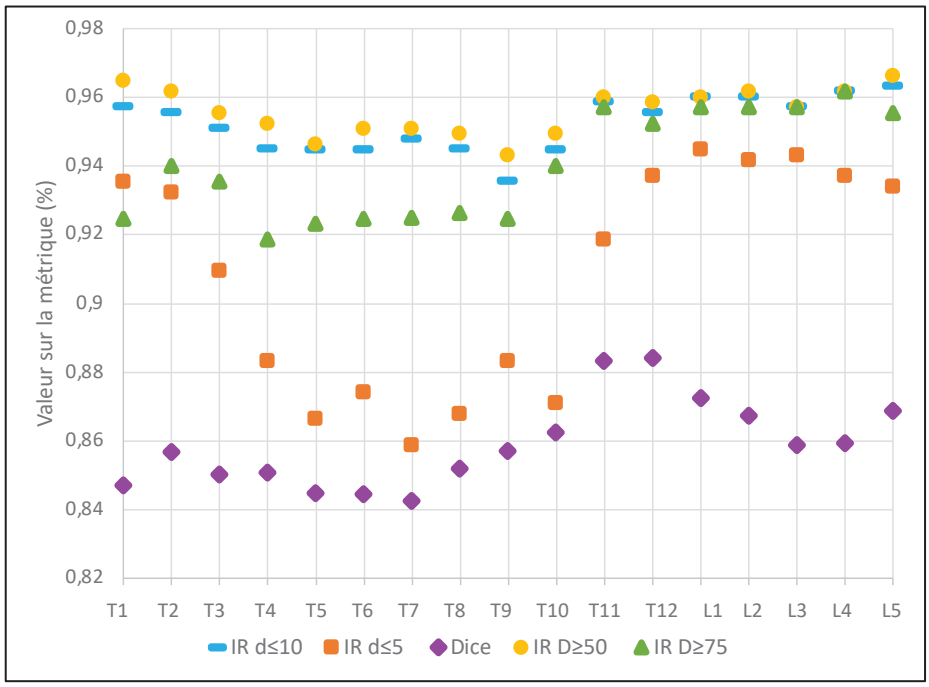



Figure 3.10 Résultats de DetectoRS de validation
selon les catégories de vertèbres

Le coefficient de Dice moyen varie entre 84,2% atteint pour T7 et 88,4% atteint pour T12. Pour les vertèbres T4 à T10, les résultats d'identification ont tendance à être plus faibles que pour les autres vertèbres.

On remarque que, sur cet intervalle de vertèbres, la baisse de résultats augmente avec la difficulté du critère d'identification. La courbe de la métrique $IR_{d\leq5}$ montre notamment des écarts de plus de 8pp entre les résultats moyens de L1 et de T7.

### 3.5.1.2 Résultats selon l'angle de Cobb

Sachant que les images radiographiques de rachis en vue frontale présentant des scolioses faibles ou extrêmes sont visuellement très différentes, on cherche à connaître l'influence de l'angle de Cobb sur les résultats. Pour cela, les distributions des résultats sur les cinq métriques en fonction de leur catégorie de scoliose sont affichées dans la Figure 3.11. Les résultats sont considérés sur les cinq plis de validation, soit 651 images.

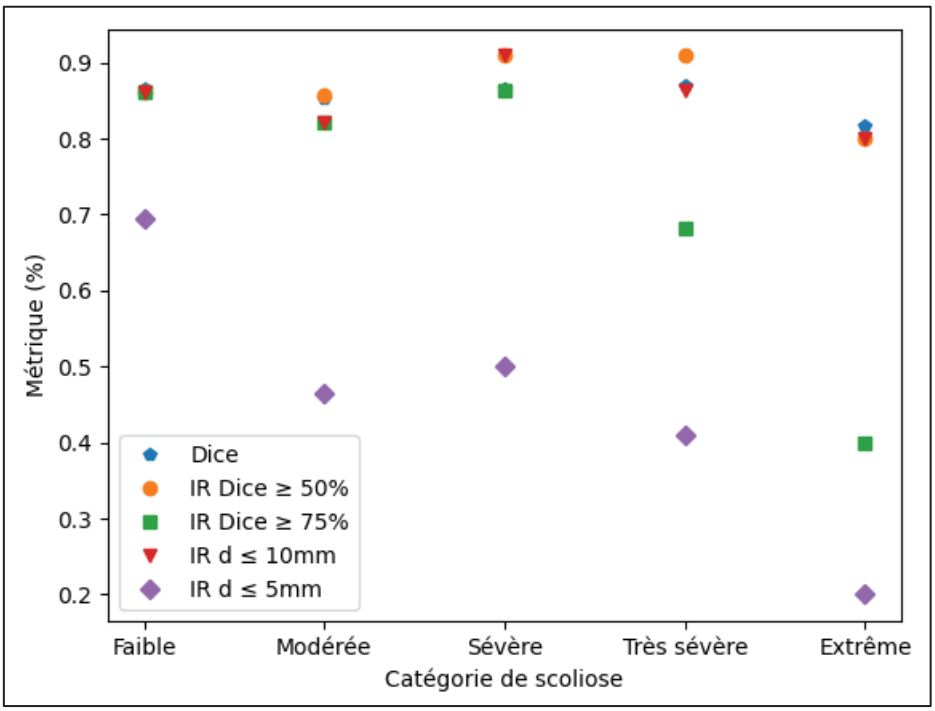



Figure 3.11 Variation des résultats moyens sur les métriques d'évaluation en fonction de la catégorie de scoliose

Pour les métriques de Dice, $IR_{D\geq 50}$ et $IR_{d\leq 10}$, la catégorie de scoliose ne semble pas avoir d'influence sur les résultats moyens. En revanche, pour $IR_{D\geq 75}$ et $IR_{d\leq}5$, métriques d'identification avec des conditions d'identifications difficiles, les résultats moyens ont tendance à chuter avec l'aggravation de la scoliose.

#### 3.5.1.3 Résultats selon les images pré- ou postopératoires

De la même manière que pour l'angle de Cobb, les métriques d'évaluation sont calculées selon deux catégories d'images : les postopératoires et les autres. Les résultats sont donnés dans le Tableau 3.6.

Tableau 3.6 Résultats sur l'ensemble de validation pour les images postopératoires et préopératoires

| | Métriques d'évaluation (%) - moyenne (médiane) | | | | |
|---|---|---|---|---|---|
| Type d'images | Dice | $IR_{D\geq50}$ | $IR_{D\geq75}$ | $IR_{d\leq10}$ | $IR_{d\leq5}$ |
| Préopératoire | 86,71 (89,52) | 96,35 (1) | 95,07 (1) | 92,53 (1) | 96,07 (1) |
| Postopératoire | 80,99 (86,66) | 90,96 (1) | 87,60 (1) | 90,07 (1) | 80,52 (88,24) |

Les images postopératoires montrent systématiquement des résultats moyens en dessous des images préopératoires. Elles présentent aussi un coefficient de Dice médian et un $IR_{d\leq5}$ plus faibles que pour les images préopératoires. La différence de résultat est particulièrement élevée pour la métrique $IR_{d\leq5}$ : le résultat des images postopératoires est 16,2% plus faible par rapport aux résultats des images préopératoires.

### 3.5.2 Distribution des résultats

Les boîtes à moustaches des résultats en validation croisée sur 5 plis sont calculées en Figure 3.12 afin de connaître la distribution des résultats par image sur les cinq métriques d'évaluation : Dice, $SR_{D\geq50}$, $SR_{D\geq75}$, $SR_{d\leq10}$ et $SR_{d\leq5}$. Sur les boîtes à moustaches, la médiane est représentée par une ligne horizontale dans la boîte, la moyenne par le triangle vert et les valeurs aberrantes (*outliers*) par les losanges noirs. Pour chacune des métriques, il existe de nombreuses valeurs aberrantes, qui correspondent à des valeurs anormalement faibles par rapport aux autres valeurs de la distribution. Soit $Q1$ le premier quartile, $Q3$ le troisième quartile et $IQR$ l'écart interquartile. Une valeur est définie comme aberrante lorsqu'elle n'appartient pas à l'intervalle $[Q1 - 1{,}5 * IRQ\ ;\ Q3\ +\ 1{,}5 * IRQ]$. Sur les 651 images de

validation, 84 sont des résultats aberrants selon la métrique de Dice, 101 selon $SR_{D\geq50}$, 149 selon $SR_{D\geq75}$, 107 selon $SR_{d\leq10}$ et 52 selon $SR_{d\leq5}$.

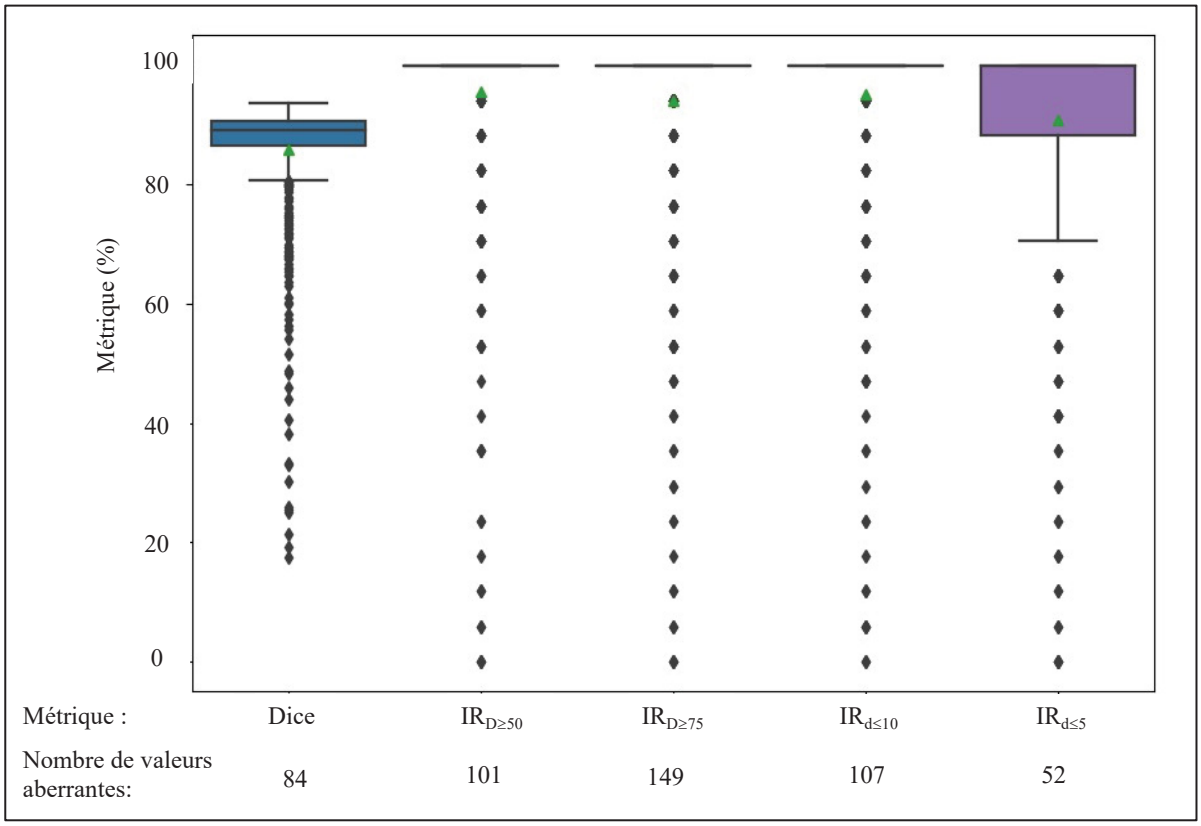



Figure 3.12 Distribution des résultats selon les cinq métriques d'évaluation. La médiane est représentée par une ligne horizontale dans la boîte, la moyenne par un triangle vert et les valeurs aberrantes par des losanges noirs

Ces nombreuses valeurs aberrantes expliquent le fait que la médiane soit systématiquement largement supérieure à la moyenne sur les métriques. Pour les métriques de taux d'identification $IR_{D\geq50}$, $IR_{D\geq75}$ et $IR_{d\leq10}$, les boîtes à moustaches sont plates : plus de 75% des images obtiennent une valeur moyenne de 100%. Sur ces métriques, les valeurs aberrantes correspondent à toutes les images dont au moins une vertèbre a mal été identifiée. $IR_{d\leq5}$, métrique de taux d'identification avec la condition la plus contraignante, montre des résultats plus distribués.

Cependant, la médiane confondue avec la borne supérieure de la boîte indique qu'au moins 50% des images obtiennent une valeur moyenne de 100%.

Les taux de succès, définis selon l'équation (3.3), sont calculés pour les quatre métriques d'identification afin connaître le nombre d'images résultantes utilisables en clinique. Ils sont affichés avec les résultats de taux d'identification moyen sur les vertèbres au Tableau 3.7.

Tableau 3.7 Taux de succès et d'identification pour DetectoRS en validation croisée

| Taux de succès SR (%) | | | |
|---|---|---|---|
| $SR_{D\geq 50}$ | $SR_{D\geq 75}$ | $SR_{d\leq 10}$ | $SR_{d\leq 5}$ |
| 84,48 | 77,11 | 83,56 | 56,68 |
| Taux d'identification des vertèbres moyens IR (%) | | | |
| $IR_{D\geq 50}$ | $IR_{D\geq 75}$ | $IR_{d\leq 10}$ | $IR_{d\leq 5}$ |
| 95,58 | 94,00 | 95,21 | 90,81 |

Les taux de succès et d'identification des vertèbres moyens associés aux mêmes conditions d'identification présentent de larges écarts de résultats. Par exemple, par rapport à $IR_{D\geq 50}$, $SR_{D\geq 50}$ est 11,61% plus faible et $SR_{d\leq 5}$ est 37,58% plus faible que $IR_{d\leq 5}$.

### 3.5.3 Recherche visuelle des erreurs récurrentes

Pour mieux comprendre les limites de DetectoRS, une revue visuelle des images résultantes est effectuée. Trois principaux types d'erreurs récurrentes sont observés : le manque de justesse des masques de segmentation, les masques de qualité médiocre et les prédictions manquées et superposées.

#### 3.5.3.1 Justesse des masques de segmentation

Par observation des images, on remarque que les masques n'épousent pas toujours la forme exacte des vertèbres, notamment au niveau des apophyses. La Figure 3.13 permet de visualiser les écarts de forme entre les masques prédits et les masques de vérité terrain. Les apophyses épineuses prédites ont tendance à être plus courtes qu'en réalité et les apophyses transverses prédites ne sont pas toujours bien placées et orientées.

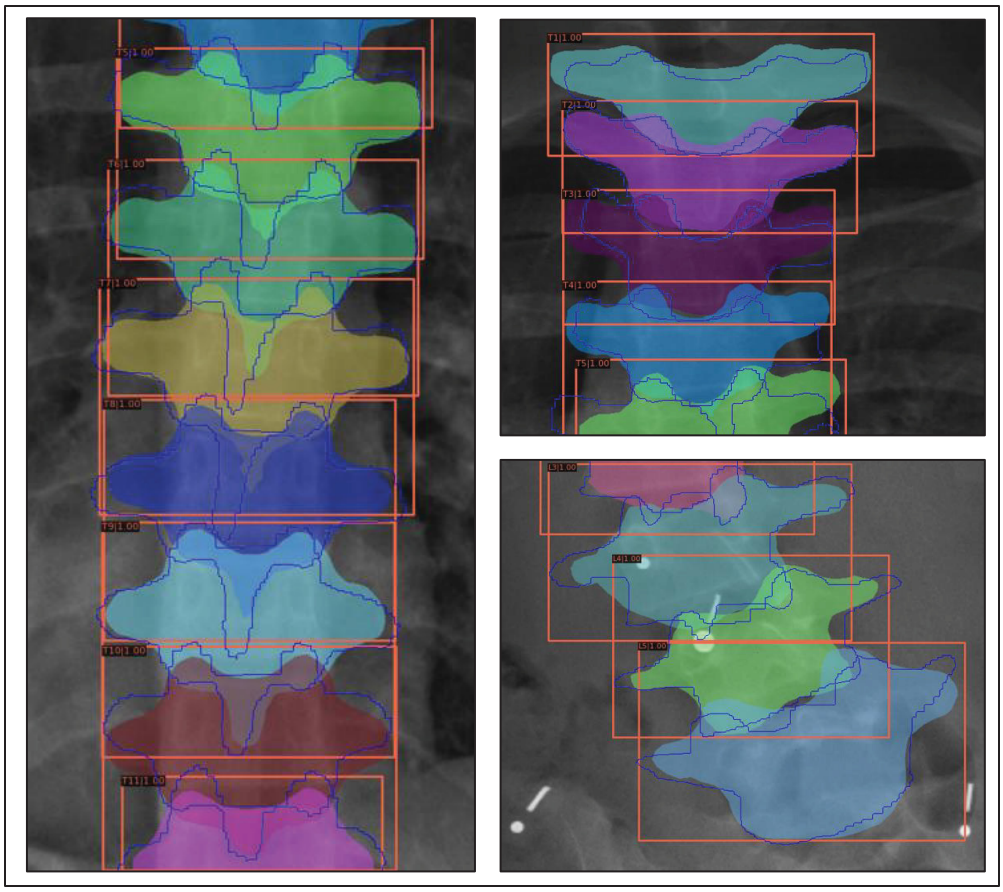


Figure 3.13 Exemples d'erreurs sur la prédiction du masque de segmentation au niveau des apophyses. Les masques remplis sont les prédictions, les contours bleus sont les masques de vérité de terrain

### 3.5.3.2 Masques de segmentation de qualité médiocre

Certaines images présentent des masques bruités, troués, difformes voire composés de plusieurs parties. La Figure 3.14 montre trois exemples d'images avec des masques prédits médiocres. En général, ces masques médiocres apparaissent sur les images de scolioses atypiques qui présentent des courbures très prononcées aux extrémités du rachis.

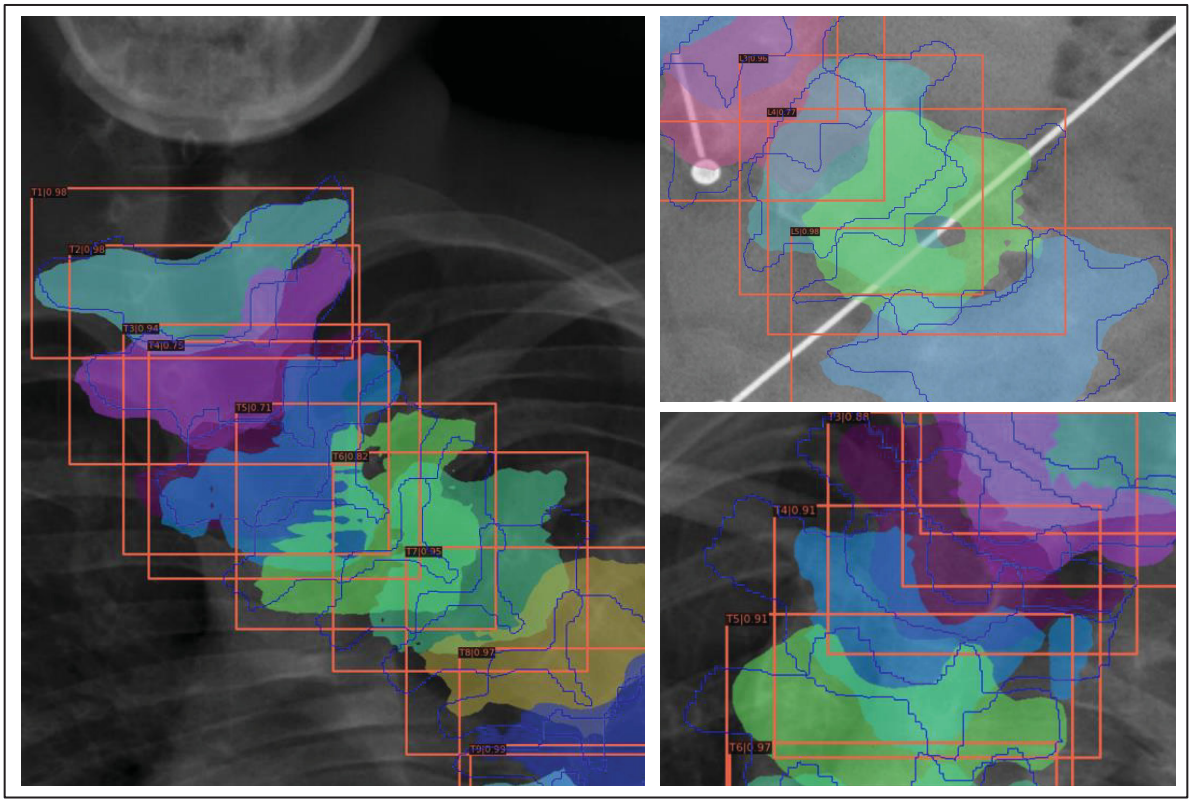


Figure 3.14 Exemples de masques de segmentation prédits médiocres. Les masques remplis sont les prédictions, les contours bleus sont les masques de vérité de terrain

### 3.5.3.3 Vertèbres manquées et prédictions superposées

Certaines images résultantes présentent des anomalies : il existe des « trous » ou des superpositions dans les prédictions. Les « trous » correspondent à des vertèbres de vérité terrain complètement manquées tandis que les superpositions correspondent à des prédictions de différentes catégories situées sur une même vertèbre de vérité terrain. La présence de

« trous » implique généralement la présence de superpositions et inversement. La Figure 3.15 illustre différents cas de prédictions manquées ou superposées. Sur l'image A., la vertèbre de vérité terrain L2 est manquée et deux prédictions se superposent sur la vertèbre de vérité terrain L1, celle de la prédiction classifiée L1 et celle de la prédiction classifiée L2. Sur l'image B., la vertèbre prédite T1 est mal localisée, elle ne correspond à aucune vertèbre de vérité terrain. La vertèbre de vérité T7 est manquée. Entre T1 et T7, toutes les prédictions sont décalées d'une catégorie. L'image C. présente aussi des erreurs en cascade. Les prédictions classifiées T12 et L1 se superposent sur la vertèbre de vérité terrain T12 et la vertèbre de vérité terrain L5 est manquée. Toutes les prédictions entre L1 et L5 sont décalées d'une catégorie.

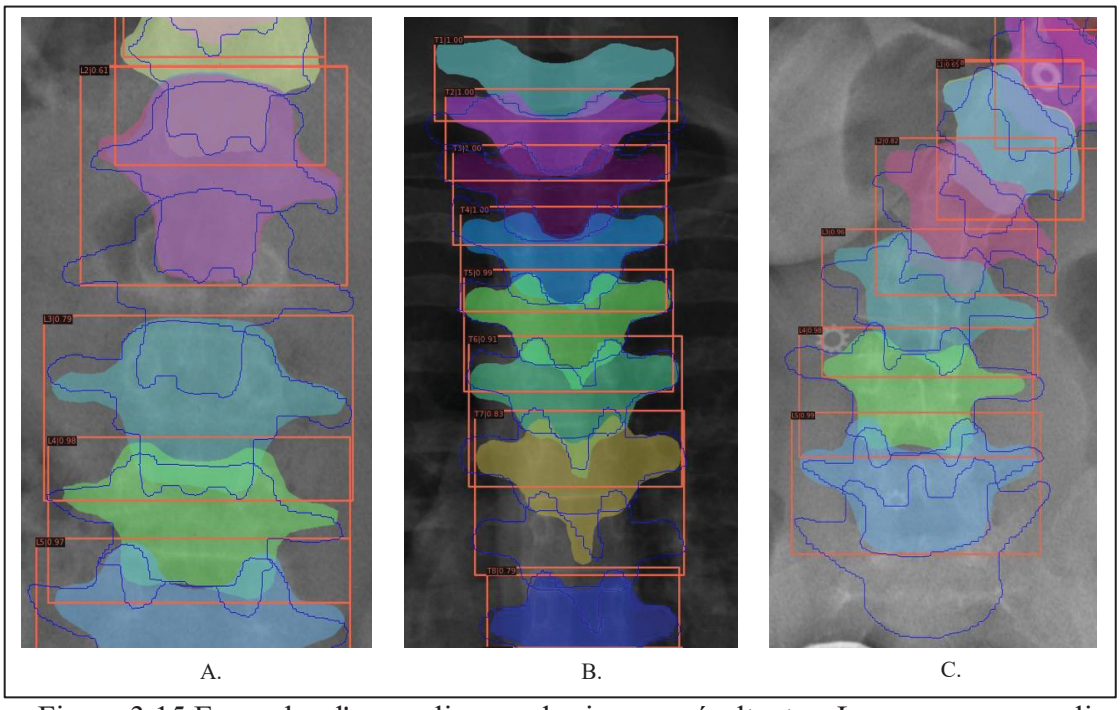



Figure 3.15 Exemples d'anomalies sur des images résultantes. Les masques remplis sont les prédictions, les contours bleus sont les masques de vérité de terrain

### 3.5.4 Erreurs sur la sélection des meilleures prédictions

Les résultats de DetectoRS après sélection des prédictions par post-traitement sont comparés à ceux où les véritables meilleures prédictions sont sélectionnées (en utilisant les données de

vérité terrain). Sur les prédictions de chaque vertèbre sur chaque image, soit 11 067 prédictions, 4,53% sont mal sélectionnées.

Dans 91,2% des cas de sélection de la mauvaise prédiction, c'est la prédiction de deuxième score le plus haut qui est véritablement la meilleure prédiction. Les erreurs de prédictions existent dans de nombreuses images résultantes : 16,28% des images de validation présentent au moins une prédiction mal choisie.

Le Tableau 3.8 rassemble les résultats sur les métriques de segmentation, de taux d'identification et de taux de succès pour les résultats en post-traitement standard et en post-traitement avec sélection des meilleures prédictions.

Tableau 3.8 Comparaison des résultats en validation obtenus avec le post-traitement et en sélectionnant les meilleures prédictions

| | Moyennes sur les métriques d'évaluation (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Dice | $IR_{D\geq50}$ | $SR_{D\geq50}$ | $IR_{D\geq75}$ | $SR_{D\geq75}$ | $IR_{d\leq10}$ | $SR_{d\leq10}$ | $IR_{d\leq5}$ | $SR_{d\leq5}$ |
| Post-traitement | 85,89 | 95,58 | 84,48 | 94,00 | 77,11 | 95,21 | 83,56 | 90,81 | 56,68 |
| Sélection des meilleures prédictions | 88,57 | 99,84 | 97,85 | 97,81 | 83,40 | 99,55 | 94,01 | 94,47 | 60,52 |
| Écart | 3,12 | 4,46 | 15,83 | 4,05 | 8,16 | 4,56 | 12,51 | 4,03 | 6,77 |

Les écarts de performances sont particulièrement marqués sur les métriques de taux d'identification et de succès, notamment lorsque la condition d'identification n'est pas très contraignante. Une sélection parfaite des prédictions permet d'augmenter $SR_{D\geq50}$ et $SR_{d\leq10}$ de respectivement 15,83% et 12,51%.

## 3.6 Discussion

### 3.6.1 *Fine-tuning* et comparaison des performances des CNNs

Les résultats obtenus sur les métriques d'évaluation par les CNNs sous différentes configurations concordent avec les allures des courbes de perte en entrainement et validation (partie 3.4.1.1). Ils témoignent tous deux d'un bon entrainement de DetectoRS et MS R-CNN. Pour RetinaMask, la courbe de perte en entrainement calculée montre que le CNN n'est pas en

sous-entrainement. L'écart prononcé entre les courbes de perte en entrainement et en validation de YOLACT s'accordent avec les performances quelque peu en retrait de YOLACT par rapport aux autres CNNs. Les entrainements divergeant sur certaines configurations sont causés par une valeur du taux d'apprentissage $\eta$ fixée trop élevée.

Par rapport aux configurations standard, les configurations optimisées des CNNs permettent chacune un léger gain de performance sur les ensembles de validation. Sur la tâche d'identification, le gain est plus élevé pour les métriques qui présentent des conditions difficiles à atteindre. Le CNN DetectoRS atteint un coefficient de Dice de 86,36% et les taux d'identification varient entre 92,09 et 95,79% en fonction de la difficulté de la condition d'identification.

En plus d'obtenir les meilleures performances sur l'ensemble de validation, DetectoRS optimisé montre sur les différentes métriques des écarts-types équivalents ou plus faibles que ceux des autres CNNs pris sous leur meilleure configuration (Tableau 3.4). L'écart-type renseigne sur les variations de performances entre les plis : un écart-type restreint correspond à un CNN stable qui généralise bien. De plus, DetectoRS se démarque particulièrement des autres CNNs sur les métriques d'identification qui demandent une précision de localisation poussée.

Les résultats des CNNs optimisés entrainés sur l'ensemble d'entrainement entier et évalués sur l'ensemble de test confirment les précédents constats des résultats en validation croisée (Tableau 3.5). Pour toutes les métriques et tous les CNNs, les résultats de test se trouvent dans les intervalles moyenne-écart-type des résultats de validation croisée en cinq plis. Cela signifie que les 4 CNNs sont capables de généraliser : ils performent de manière similaire sur les données de validation et sur des données inconnues. Comme le suggérait les résultats en validation, DetectoRS est le CNN le plus performant et YOLACT est en retrait. DetectoRS devance les autres CNNs sur toutes les métriques. De la même manière que pour les résultats en validation, DetectoRS creuse l'écart de performance sur les métriques d'identification avec des conditions exigeantes. On remarque aussi que DetectoRS est le seul CNN à présenter des résultats en test au-dessus de la moyenne de ses résultats en validation : il atteint un Dice de

86,36% en test et un Dice moyen de 85,89% en validation. Cela peut s'expliquer par le fait qu'en test, plus de données d'entrainement ont été utilisées qu'en validation. DetectoRS apprend mieux avec les images supplémentaires.

Les résultats des analyses des intervalles de Bonferroni (partie 3.4.2) confirment bien que YOLACT est moins performant que les autres CNNs. En revanche, la supériorité de DetectoRS n'est prouvée que pour la métrique d'identification la plus contraignante, $IR_{d\leq5}$. Ce résultat peut s'expliquer par le fait que les intervalles de Bonferroni sont des tests relativement prudents, utilisant de larges intervalles de confiance (McDonald, 2015).
Ces résultats, obtenus sur des images radiographiques, sont en accords avec les résultats obtenus sur les images optique de l'ensemble COCO test-dev (donnés au Tableau 3.2) : DetectoRS se place en première position, MS R-CNN et RetinaMask montrent des performances similaires tandis que YOLACT prend la dernière position. La supériorité de DetectoRS par rapport aux autres CNNs semble se transmettre aux images radiographiques. Alors que les résultats de MS R-CNN sont semblables à ceux obtenus par (Kónya et al., 2021) avec Mask R-CNN, YOLACT performe nettement moins bien dans cette étude que dans celle de (Kónya et al., 2021). La forme plus compliquée des vertèbres ainsi que la superposition accrue de ces dernières peuvent être à l'origine de moins bon apprentissage de YOLACT, qui produit ainsi des masques moins précis.

### 3.6.2 Analyses des résultats de DetectoRS

#### 3.6.2.1 Influence des données

Les vertèbres situées aux extrémités du rachis sont en moyenne mieux identifiées que celles situées au centre du rachis (Figure 3.10). Plus la métrique possède une condition d'identification difficile, plus l'écart de résultat d'identification entre les vertèbres au centre et aux extrémités est marqué. Cela signifie que DetectoRS a du mal à localiser les vertèbres situées au centre du rachis de manière très précise. La difficulté de détection des apophyses

peut être responsable des localisations imprécises. En effet, les masques de segmentation de vérité terrain des vertèbres thoraciques sont particulièrement marqués par la forme des apophyses : les apophyses épineuses des thoraciques présentent une orientation ainsi qu'une longueur qui les rend très apparentes en vue frontale.

La précision de localisation se dégrade aussi quand l'angle de Cobb devient très grand sur une image (Figure 3.11). Ce constat peut s'expliquer par les écarts d'apparence entre les images appartenant à un même type de scoliose. Les scolioses de type « Faible » partagent une apparence très similaire : le rachis en vue frontale se rapproche d'une droite verticale. Au contraire, les scolioses très prononcées peuvent prendre des apparences multiples : les courbures du rachis sont plus ou moins nombreuses, localisées à différents endroits et sont plus ou moins prononcées. Leur différence d'apparence les rend plus difficiles à traiter par le CNN. De plus, plus l'angle de Cobb est grand, plus les vertèbres ont tendance à être très superposées, rendant la tâche de segmentation et d'identification plus complexe. Enfin, on note aussi que le nombre d'images appartenant à une catégorie de scoliose diminue avec la sévérité de cette dernière. Le nombre restreint de cas de scolioses très sévères et extrêmes est une limitation supplémentaire pour leur apprentissage.

Au-delà du problème d'occlusion lié à la présence d'outils d'instrumentation, les résultats en retrait des images postopératoires (Tableau 3.6) peuvent aussi être expliqués par le fait que ces images représentent seulement 13,7% de la base de données. De plus, leur répartition dans les différents ensembles d'entrainement, de validation et de test n'a pas été examinée.

#### 3.6.2.2 Limites de DetectoRS

Bien que les métriques d'identification présentent des moyennes relativement élevées, les métriques de taux de succès montrent des résultats moins satisfaisants (Tableau 3.7). Le taux de succès varie entre 84,48% pour $IR_{D\geq 50}$ et 56,68% pour $IR_{d\leq 5}$. Cela signifie que de nombreuses images résultantes possèdent au moins une vertèbre mal identifiée et ne sont donc pas utilisables en clinique.

L'analyse visuelle des images résultantes mène à l'observation de trois types d'erreurs :

1) Un manque de justesse sur les masques
2) Des masques de qualité médiocre
3) Des anomalies de « trous » et superposition de masques

Comme l'indiquent les résultats de Dice, les masques de segmentation manquent parfois de justesse. L'erreur sur la justesse est tout de même acceptable quand elle ne dépasse pas un seuil de Dice minimal qui rendrait l'identification de la vertèbre nulle.

Bien que cette erreur de justesse sur les masques soit généralement maitrisée, il arrive que les masques prédits soient de qualité médiocre. Ils sont anatomiquement absurdes. Ces mauvaises prédictions témoignent d'une faible capacité de généralisation de DetectoRS pour les cas de scoliose atypiques. DetectoRS n'a pas suffisamment appris ce type de scoliose et n'est pas capable de les traiter correctement.

Les anomalies de « trous » et superpositions de prédictions visibles dans certaines images résultantes sont des erreurs de classification. Par exemple, sur l'image A. de la Figure 3.15, la prédiction classifiée L2 consiste en réalité à une prédiction L1. En fait, les catégories attribuées à des prédictions spatialement voisines sont parfois confondues. Ces anomalies sont particulièrement contrariantes, car elles correspondent à des prédictions complètement fausses : l'identification est incorrecte et le score de Dice est quasiment nul. De plus, ces anomalies se produisent parfois en cascade sur une image. L'image présente alors des scores d'évaluation très faibles. Ces images, sur lesquelles au moins une vertèbre est mal identifiée, sont en « échec ». Elles correspondent ainsi en grande partie aux valeurs aberrantes des distributions visualisées dans la

Figure 3.12 et sont largement responsables des faibles taux de succès. On remarque aussi que les vertèbres manquées et les prédictions superposées dans une image ne semblent pas dépendre de la qualité des autres prédictions dans l'image. Ces anomalies existent sur les images de scolioses de toutes catégories.

Malgré quelques erreurs, le score de confiance reste donc un assez bon indicateur de la qualité d'une prédiction. Cependant, les mauvais choix de prédictions exercent une influence importante au niveau des performances, notamment sur les taux de succès. En effet, une

sélection parfaite des prédictions permet d'atteindre des taux de succès impressionnants (Tableau 3.8). Les prédictions de très bonne qualité existent, elles ne correspondent juste pas toujours au score de confiance le plus élevé et, de ce fait, ne sont pas choisies par le post-traitement.

Le problème de sélection des prédictions est directement lié au phénomène des anomalies sur les images. La sélection d'une prédiction mal classifiée est responsable des anomalies de « trous » et superpositions sur les images. Quand les prédictions sont parfaitement classifiées, ces anomalies disparaissent et les images passent du statut « échec » à « succès », faisant largement augmenter le taux de succès.

## 3.7 Conclusion

Ce chapitre a permis de vérifier que les CNNs de segmentation d'instance, développés pour fonctionner sur des images naturelles, sont aussi capables de fonctionner sur des images radiographiques EOS frontales. DetectoRS, MS-RCNN, RetinaMask et YOLACT sont appliqués aux tâches de segmentation et d'identification individuelle des vertèbres sur des radiographies EOS de rachis atteints d'AIS. Le choix des métriques d'évaluation fait l'objet d'un effort particulier. Les métriques sont interprétables dans le domaine médical et sont pertinentes par rapport à la base de données. Une opération de *fine-tuning* est réalisé pour obtenir la configuration optimale de chacun des quatre CNNs. DetectoRS obtient les meilleures performances sur toutes les métriques d'évaluation. Sur l'ensemble de test, il atteint un coefficient de Dice de 86,6% et un taux d'identification moyenné sur les quatre métriques d'identification de 94,83% (Tableau 3.5). Les analyses statistiques comparatives confirment simplement que DetectoRS est meilleur que les trois autres CNNs sur la métrique de taux d'identification la plus contraignante.

DetectoRS est sélectionné pour la suite de l'étude. L'analyse approfondie de ses résultats montre que des erreurs dans la sélection des prédictions causée par de multiples facteurs décrits précédemment à la sortie de DetectoRS impactent particulièrement les performances et plus spécifiquement le taux d'images utilisable en milieu clinique.

# CHAPITRE 4

# PROPOSITION D'UNE VERSION DE DETECTORS SPÉCIALISÉE À LA SEGMENTATION ET L'IDENTIFICATION DES VERTÈBRES DANS LES RADIOGRAPHIES FRONTALES EOS : V-DETECTORS

## 4.1 Introduction

À l'issue de l'étude comparative effectuée dans le chapitre précédent, le CNN de segmentation d'instance DetectoRS est sélectionné. Ce dernier performe bien : DetectoRS atteint un coefficient de Dice de 86,6% et un taux d'identification de 94,88% pour la condition d'identification Dice ≥ 75%.

La revue de littérature montre que les méthodes CNN pensées pour la segmentation et la détection des vertèbres sur des images radiologiques mettent en place différentes stratégies pour insérer des connaissances préalables relatives au rachis supplémentaires dans le processus (Lessmann *et al.*, 2019 ; Payer *et al.*, 2021 ; Sekuboyina *et al.*, 2021 ; Kim *et al.*, 2021). Cette spécialisation à la tâche permet d'améliorer les performances, notamment en produisant des résultats anatomiquement plausibles.

L'objectif de ce chapitre est d'améliorer les performances de DetectoRS pour la tâche de segmentation et d'identification des vertèbres sur des radiographies frontales EOS en le spécialisant à cette tâche.

La spécialisation de DetectoRS résulte en la mise en place et l'évaluation du CNN V-Detectors, qui est une version de DetectoRS intégrant différentes stratégies d'améliorations. Les propositions d'amélioration étudiées, développées à partir de la littérature et de l'analyse des résultats de DetectoRS, sont basées sur l'intégration de connaissances préalables supplémentaires dans la méthode. Elles consistent en l'ajoute d'un pré-entrainement ajusté au type d'images et à l'utilisation d'un post-traitement et d'une fonction de perte anatomiquement contraignants.

Les propositions sont d'abord évaluées sur l'ensemble d'entrainement en utilisant la stratégie de validation croisée à cinq plis. V-DetectoRS, version de DetectoRS qui intègre les propositions évaluées comme bénéfiques, est finalement évalué sur l'ensemble de test.

## 4.2 Méthodologie

La mise en place d'une version améliorée de DetectoRS passe par plusieurs étapes :

- L'identification de pistes d'amélioration compte tenu de la revue de littérature et de l'analyse des limites de DetectoRS ;
- La conception et l'intégration des stratégies d'amélioration à DetectoRS ;
- L'évaluation des stratégies de spécialisation en validation croisée à cinq plis. À l'issue, les stratégies sont jugées bénéfiques ou non ;
- L'évaluation de V-DetectoRS, intégrant les stratégies de spécialisation bénéfiques, sur l'ensemble de test.

### 4.2.1 Identification des stratégies de spécialisation

Dans l'étude menée au chapitre précédent, les *backbones* des CNNs sont utilisés pré-entrainés sur l'ensemble ImageNet. Cependant, la revue de littérature souligne le fait que le pré-entrainement réalisé sur ImageNet n'est pas adapté à l'utilisation d'images radiographiques. Les images radiographiques ayant une apparence très éloignée des images optiques, le transfert des caractéristiques apprises est difficile (Raghu *et al.*, 2019). L'utilisation d'ensembles de données cibles et sources appartenant au même domaine rend le pré-entrainement plus efficace et semble même permettre des gains de performances (Romero *et al.*, 2020). Un pré-entrainement ajusté aux images radiographiques de rachis apparaît comme une première stratégie de spécialisation.

La littérature met aussi en avant l'intérêt d'intégrer des connaissances préalables relatives au contexte de manière explicite dans le processus. Les méthodes CNN spécialisées pour la tâche d'identification et de segmentation de vertèbres utilisent toutes des informations explicites

relatives à l'anatomie du rachis, notamment par rapport à l'ordre fixé des vertèbres dans le rachis. L'ajout de contraintes anatomiques dans le processus d'identification et de segmentation permet d'écarter les résultats anatomiquement invraisemblables. L'analyse des limites de DetectoRS, effectuée dans la partie 3.6.2.2 du CHAPITRE 3, montre justement que certains résultats sont anatomiquement impossibles : des images résultantes présentent des « trous » et des superpositions dans les détections. Ces anomalies, issues de la sélection de prédictions dont l'identification est erronée, rendent les images résultantes inutilisables dans un contexte clinique. Elles expliquent les valeurs restreintes de taux de succès par rapport aux taux d'identification (Tableau 3.7).

L'analyse montre tout de même que des prédictions bien localisées et identifiées existent quasiment systématiquement, elles ne sont juste pas sélectionnées lors du post-traitement (Tableau 3.8). La mise en place d'un post-traitement anatomiquement contraignant semble ainsi pertinente. La sélection des prédictions ne dépend alors pas seulement du score de confiance prédit par DetectoRS, mais aussi du respect de contraintes d'ordre anatomique.

La stratégie du post-traitement anatomiquement contraignant présente tout de même quelques limites : il suppose l'utilisation d'intervalles statistiques dont la fiabilité peut être mise en question et est dépendant des prédictions en sortie de DetectoRS. Les connaissances préalables relatives à l'arrangement des vertèbres dans le rachis sont seulement utilisées en sortie du CNN, elles ne sont pas intégrées à l'apprentissage. À la manière de (Kervadec *et al.*, 2019), il semble intéressant d'intégrer directement des contraintes anatomiques sous la forme de pénalité dans la fonction de perte de DetectoRS. L'objectif est que DetectoRS apprenne à produire des prédictions anatomiquement possibles.

Plus spécifiquement, pour le post-traitement et le terme de pénalité, les contributions techniques consistent en la modélisation des distances intervertébrales, respectivement dans le processus de sélection des prédictions et le processus d'identification et de localisation des vertèbres.

### 4.2.2 Principe des stratégies de spécialisation

Trois propositions de spécialisation de DetectoRS sont mises en place : un pré-entrainement ajusté au domaine, un post-traitement et une fonction de perte de régression tenant compte des distances intervertébrales. Le principe de chacune d'elles est expliqué ci-dessous.

#### 4.2.2.1 Pré-entrainement ajusté

Un pré-entrainement plus ajusté aux images radiographiques utilisées dans cette étude est mis en place. Il existe deux principaux ensembles de données de radiographies frontales de la région du rachis publiquement disponibles : CheXpert (Irvin *et al.*, 2019) et ChestX-Ray (Wang *et al.*, 2017). Les deux ensembles sont similaires : ils regroupent des milliers d'images radiographiques labellisées de thorax. Les labels renseignent sur la présence de différentes maladies pulmonaires dans les images. ChestX-Ray est préférée à CheXpert, car elle contient des images de meilleure résolution – 1024×1024 contre seulement 550×550 pour CheXpert. Le pré-entrainement est effectué sur la version 14 de ChestX-ray, ChestX-ray14. ChestX-ray14 contient 51 649 images de radiographies frontales de thorax labellisées selon 14 catégories de maladies. Certaines images présentent plusieurs maladies.

ResNet50 est ainsi entrainé pour la tâche de classification des images ChestX-ray14. Le code de prétraitement, d'entrainement et d'évaluation est basé sur le dépôt GitHub de Paloukari (Garyfallos, Biseda et Khan, 2019). Les données d'entrainement et de test sont réparties selon le rapport 85:15.

Le *backbone* Resnet50 entrainé sur ChestX-ray14 est utilisé en tant que pré-entrainement de DetectoRS sous deux configurations :

- Les poids transférés ne sont fixés que sur les couches les plus basses;
- Aucun des poids transférés n'est fixé, le pré-entrainement est utilisé en tant qu'initialisation des poids.

DetectoRS est entrainé sous les deux configurations de pré-entrainement de la même manière et sur les mêmes ensembles que pour le CHAPITRE 3. La stratégie de validation croisée en 5 plis est réutilisée.

#### 4.2.2.2 Post-traitement avec respect de la structure anatomique

Un nouveau post-traitement, tirant partie des connaissances géométriques connues sur le rachis, est mis en place. Ce post-traitement prend en compte les distances intervertébrales verticales des prédictions pour détecter les images résultantes présentant des anomalies. Ces résultats anormaux sont alors corrigés en sélectionnant les prédictions qui garantissent au mieux le respect des distances intervertébrales. La distance intervertébrale est calculée entre les centres de deux masques de segmentation des vertèbres adjacentes, comme illustré sur la Figure 4.1. Pour une image, 16 distances intervertébrales sont donc calculées.

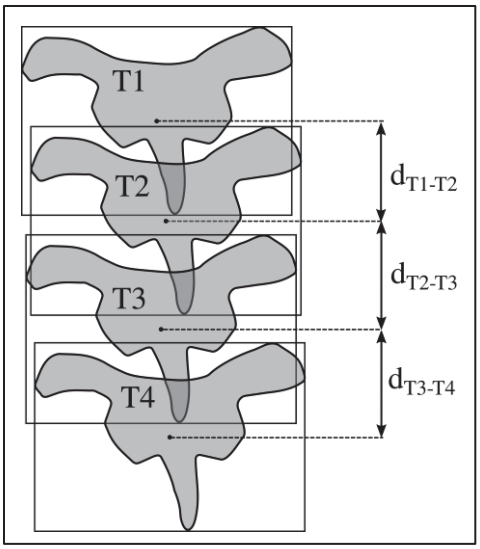



Figure 4.1 Calcul des distances verticales intervertébrales

Afin de détecter les anomalies et corriger les résultats, des intervalles statistiques pour chaque distance intervertébrale sont établis. Les 16 distributions des distances intervertébrales, propres à chaque vertèbre, sont calculées sur l'ensemble d'entrainement. Sachant que la moitié des distributions suivent une loi normale et que le nombre d'échantillons est grand, on décide

d'utiliser un intervalle de tolérance de la moyenne basé sur la loi normale. L'intervalle de tolérance à 95% $[\bar{m} - 2\sigma\ ;\ \bar{m} + 2\sigma\ ]$ avec $\bar{m}$ la moyenne et $\sigma$ l'écart-type est choisi.

Dans une première étape, le post-traitement attribue les prédictions et repère les images anormales. Le post-traitement sélectionne pour chaque image résultante les prédictions selon le score de confiance le plus élevée pour chaque vertèbre et vérifie que chacune des distances intervertébrales résultantes est contenue dans l'intervalle de tolérance correspondant. Si une image résultante ne respecte pas un des intervalles de confiance, l'image est catégorisée comme anormale.

Dans une deuxième étape, le post-traitement effectue la correction des images détectées comme anormales. Les prédictions de score de confiance le plus élevé et de score de confiance le deuxième plus élevé sont considérées. Le post-traitement sélectionne la combinaison de prédictions optimales $C_{opt}$ parmi les $2^{17}$ combinaisons possibles avec la fonction de score $S$ donnée en équation (4.1). La fonction de score $S$ indique l'intensité avec laquelle une combinaison de prédictions $j$ respecte ou non les intervalles de tolérance. Soit $D_j$ le vecteur des 16 distances intervertébrales $d_{j,i}$ des prédictions d'une combinaison $j$ testée. Pour chaque combinaison, la fonction $S$ calcule un score à partir des intervalles selon la formule donnée dans l'équation (4.1).

$$S(D_j) = \sum_{i=1}^{16} s(d_{j,i}) \text{ où } s(d_{j,i}) = \begin{cases} \mathrm{a}_i - d_{j,i}, & \text{si } d_{j,i} < \mathrm{a}_i \\ d_{j,i} - \mathrm{b}_i, & \text{si } d_{j,i} > \mathrm{b}_i \\ 0 & \text{sinon} \end{cases} \tag{4.1}$$

avec $[\mathrm{a}_i;\ \mathrm{b}_i]$ intervalle de confiance pour la distance intervertébrale $i$ .

La combinaison optimale $C_{opt}$, définie selon l'équation (4.2), est celle qui atteint le score minimal de $S$, soit celle qui respecte au mieux les contraintes de distances intervertébrales.

$$C_{opt} = \underset{j}{\operatorname{argmin}}\, S(D_j) \tag{4.2}$$

Le post-traitement révise seulement les images qui présentent des distances intervertébrales jugées anormales avec les intervalles de confiance. De ce fait, on diminue le risque que le post-traitement détériore les résultats des images anatomiquement correctes.

Pour qualifier l'efficacité du post-traitement sur les images détectées comme anormales, on calcule le nombre d'images dégradées et d'images améliorées pour chaque métrique de taux de succès. L'efficacité correspond au nombre d'images améliorées sur le nombre d'images modifiées. Les images améliorées correspondent aux images qui étaient au statut « échec » et qui passent au statut « succès » après application du post-traitement. Au contraire, les images dégradées désignent les images qui étaient au statut « succès » et qui passent au statut « échec » après le post-traitement.

Ces images dégradées correspondent à des cas où la condition d'identification est respectée alors que l'intervalle de confiance de distance intervertébrale ne l'est pas. Les images dégradées sont dues à des cas particuliers présentant des distances intervertébrales de vérité terrain en dehors des intervalles de confiance ou à une trop grande divergence entre les métriques d'identification et le respect de l'intervalle de tolérance (par exemple, $IR_{D\geq 50}$ est respecté pour deux prédictions voisines, mais la distance entre leurs centres est très élevée).

#### 4.2.2.3 Fonction de perte de régression avec respect de la structure anatomique

En se basant sur le travail de Kervadec (Kervadec *et al.*, 2019), les connaissances préalables sont intégrées dans une fonction de perte de DetectoRS sous la forme de pénalités. La fonction de perte doit pénaliser le CNN sur les prédictions qui montrent des distances intervertébrales statistiquement anormales. Le calcul des distances intervertébrales se fait sur l'axe vertical, car ce dernier traduit l'arrangement fixe que suit la structure anatomique du rachis, même scoliotique.

La fonction de perte doit mettre des contraintes entre les prédictions de catégories voisines. Le but est de contraindre les positions des prédictions adjacentes entre elles. Le fait que les contraintes soient mises en place entre les prédictions rend la mise en place de la fonction de perte plus complexe.

Les contraintes entre les prédictions se traduisant en termes de distances à respecter, la fonction de perte de pénalité $Loss_{InterV}$ mise en œuvre est ajoutée à la fonction de perte de régression des boîtes englobantes $Loss_{SmoothL1}$. La fonction de perte $Loss_{SmoothL1}$ correspond à la fonction de régression des boîtes englobantes de DetectoRS, *Smooth L1 Loss*, définie selon l'équation (4.3). Le principe de fonctionnement de $Loss_{SmoothL1}$ est rappelé avant de présenter la fonction $Loss_{InterV}$.

$Loss_{SmoothL1}$ prend en entrée :

- Le tenseur $Pred$ des boîtes englobantes estimées qui ne correspondent pas à l'arrière-plan. $Pred$ est de dimension $(P, 4)$ avec $P$ le nombre de prédictions et 4 les coordonnées $[x_1, y_1, x_2, y_2]$ de la boîte englobante ( $x_1, y_1$ coordonnées du coin haut gauche et $x_2, y_2$ coordonnées du coin bas droit).
- Le tenseur $Cible$ des boîtes englobantes de vérité terrain correspondant aux véritables catégories des boîtes englobantes estimées, non pas des catégories estimées. $Cible$ est de dimension $(P, 4)$ avec $P$ le nombre et 4 les coordonnées $[x_1, y_1, x_2, y_2]$ des boîtes englobantes de vérité terrain correspondant aux boîtes englobantes prédites.

La perte $Loss_{SmoothL1}$ se calcule à partir des différences entre les tenseurs $Pred$ et $Cible$. Soit $pred_{i,j}$ un élément de $Pred$ correspondant à la prédiction $i$ et à la coordonnée $j$ et $cible_{i,j}$ l'élément de $Cible$ correspondant à $pred_{i,j}$. $Loss_{SmoothL1}$ est donné par l'équation (4.3) (Girshick, 2015a).

$$Loss_{SmoothL1} = \sum_{i=1}^{P} \sum_{j \in \{x_1, y_1, x_2, y_2\}} SmoothL1\,(pred_{i,j} - cible_{i,j})$$

$$\text{où} \quad SmoothL1\,(x) = \begin{cases} \frac{1}{2}\,x^2\,, & \text{si } |x| < 1 \\ |x| - \frac{1}{2}\,, & \text{sinon} \end{cases} \tag{4.3}$$

Le tenseur $Loss_{SmoothL1}$ est moyenné pour obtenir le terme de perte en régression.

L'objectif de la fonction $Loss_{InterV}$ est de pénaliser le réseau quand les prédictions présentent des distances intervertébrales statistiquement invraisemblables sur l'axe vertical. La pénalité est donc appliquée seulement quand les distances prédites se situent hors d'un intervalle de tolérance établi à partir des distances de vérité terrain. Contrairement au prétraitement, il est possible de se servir des distances intervertébrales de vérité terrain. L'utilisation des distances intervertébrales de vérité terrain, calculées à partir des boîtes englobantes de vérité terrain, permet de définir des intervalles de tolérances plus justes. Soit $dist_{vérité}$, la distance intervertébrale de vérité terrain relative à deux vertèbres adjacentes sur une image. L'intervalle de tolérance est fixé à $[0{,}8 * dist_{vérité}; 1{,}2 * dist_{vérité}]$.

La fonction $Loss_{InterV}$ s'intéresse aux distances entre les prédictions de catégories voisines. La perte est calculée prédiction par prédiction, en considérant ses prédictions voisines « précédentes » et « suivantes » : quand on regarde une prédiction de catégorie T3, on s'intéresse à la distance entre cette prédiction et les prédictions de catégories T2 (vertèbre « précédente ») et à la distance entre cette prédiction et les prédictions de catégories T4 (vertèbre « suivante »). Il arrive qu'il n'y ait aucune prédiction de catégorie suivante ou précédente, notamment au début de l'entrainement. Dans ce cas, il est nécessaire d'appliquer une perte pour marquer cette erreur : une perte fixée à 1 est attribuée. Quand des prédictions de catégories T1 ou L5 sont considérées, les pertes liées aux prédictions respectivement précédentes et suivantes sont nulles.

Le principe de fonctionnement de $Loss_{InterV}$ est expliqué ci-dessous.

$Loss_{InterV}$ prend en entrée :

- Le tenseur $Pred$ des boîtes englobantes estimées qui ne correspondent pas à l'arrière-plan. $Pred$ est de dimension $(P, 4)$ avec $P$ le nombre de prédictions et 4 les coordonnées $[x_1, y_1, x_2, y_2]$ de la boîte englobante.
- Le tenseur $Cible$ des boîtes englobantes de vérité terrain correspondant aux véritables catégories des boîtes englobantes estimées. $Cible$ est de dimension $(P, 4)$ avec $P$ le nombre et 4 les coordonnées $[x_1, y_1, x_2, y_2]$ des boîtes englobantes de vérité terrain correspondant aux boîtes englobantes prédites.

- Le tenseur $Label$ des catégories de vérité terrain correspondantes aux boîtes englobantes du tenseur $Pred$. $Label$ est de dimension $(P, 1)$ avec $P$ le nombre de prédictions.

$Pred$ , $Cible$ et $Label$ sont classés selon l'ordre croissant des catégories, soit de T1 à L5.
Soit $pred$ une prédiction considérée dans le tenseur $Pred$ :

- $preds_{suiv,i}$ désigne une prédiction $i$ de catégorie suivante parmi le nombre total $N_{suiv}$ de prédictions de catégorie suivante.
- $preds_{pré,i}$ désigne une prédiction $i$ de catégorie précédente parmi le nombre total $N_{pré}$ de prédictions de catégorie précédente.
- $cible$ est la vérité terrain correspondante à $pred$, $cible_{pré}$ est la vérité terrain correspondante aux $preds_{pré,i}$ et $cible_{suiv}$ est la vérité terrain correspondante aux $preds_{suiv,i}$.

La distance sur l'axe vertical entre une $pred$ et une $pred_{suiv,i}$ est définie par :

$$dist_{suiv,i} = Dist_Y\left(pred, pred_{suiv,i}\right) \tag{4.4}$$

La distance sur l'axe vertical entre une $pred$ et une $pred_{pré,i}$ est définie par :

$$dist_{pré,i} = Dist_Y\left(pred_{pré,i}, pred\right) \tag{4.5}$$

La distance sur l'axe vertical entre une $cible$ et sa $cible_{pré}$ est définie par :

$$dist_{vérité,pré} = Dist_Y(cible_{pré}, cible) \tag{4.6}$$

La distance sur l'axe vertical entre une $cible$ et sa $cible_{suiv}$ est définie par :

$$dist_{vérité,suiv} = Dist_Y(cible, cible_{suiv}) \tag{4.7}$$

où $Dist_Y(A, B) = \frac{1}{2}\left((y_{1B} + y_{2B}) - (y_{1A} + y_{2A})\right)$

Pour $pred$ in $Pred$ :

$$Loss_{pred,pré} = \begin{cases} \frac{1}{N_{pré}} \sum_{i=1}^{N_{pré}} loss(dist_{pré,i}, dist_{vérité,pré}) & \text{si } Label_i > 0 \text{ et } pred_{pré,i} \neq 0 \\ 1 & \text{si } pred_{pré,i} = 0 \\ 0 & \text{si } Label_i = 0 \end{cases} \quad (4.8)$$

$$Loss_{pred,suiv} = \begin{cases} \frac{1}{N_{suiv}} \sum_{i=1}^{N_{suiv}} loss(dist_{suiv,i}, dist_{vérité,suiv}) & \text{si } Label_i < L5 \text{ et } pred_{suiv,i} \neq 0 \\ 1 & \text{si } pred_{suiv,i} = 0 \\ 0 & \text{si } Label_i = L5 \end{cases} \quad (4.9)$$

où $loss(dist, dist_{vérité}) = \begin{cases} (a - dist) & \text{si } dist < a \\ (dist - b) & \text{si } dist > b \ avec \ [a, b] = [0{,}8 * dist_{vérité} \ ; 1{,}2 * dist_{vérité}] \\ 0 & \text{sinon} \end{cases}$

$$Loss_{InterV} = \sum_{pred \ in \ Pred} Loss_{pred,pré} + Loss_{pred,suiv} \quad (4.10)$$

$Loss = Loss_{Smooth\ L1} + \lambda Loss_{InterV}$ est moyenné pour obtenir un terme unique de perte en régression. $\lambda \in \mathbb{R}$ est un hyperparamètre à estimer contrôlant la pondération du terme de pénalité $Loss_{InterV}$.

Les paramètres sont actualisés selon l'algorithme de descente de gradient :

$$y' = y - \eta \left( \frac{\partial Loss}{\partial y} \right) \quad (4.11)$$

On s'assure que la perte $Loss_{InterV}$ est bien dérivable. Le gradient de $Loss_{pred,suiv}$ est calculé sur $y$ en équation (4.12).

$$\frac{\partial Loss_{pred,suiv}}{\partial y} = \begin{cases} \frac{1}{N_{suiv}} \sum_{i=1}^{N_{suiv}} \frac{\partial loss\left(dist_{suiv,i}, dist_{vérité,suiv}\right)}{\partial y} & \text{si } Label_i < L5 \text{ et } pred_{suiv,i} \neq 0 \\ 0 & \text{si } pred_{suiv,i} = 0 \\ 0 & \text{si } Label_i = L5 \end{cases} \quad (4.12)$$

$$\text{avec } \frac{\partial\left(dist_{suiv,i}, dist_{vérité,suiv}\right)}{\partial y} = \begin{cases} 0{,}5 & \text{si } dist < a \\ -0{,}5 & \text{si } dist > b \\ 0 & \text{sinon} \end{cases}$$

$$o\grave{u}\ [a,b] = [0{,}8 * dist_{vérité}\ ; 1{,}2 * dist_{vérité}]$$

Physiquement, le gradient a du sens. Dans le cas A. de la Figure 4.2, $\frac{\partial Loss_{pred,suiv}}{\partial y} < 0$, $y'$ tend vers une valeur plus grande. Il sera donc situé plus bas dans l'image. Dans le cas B., $\frac{\partial Loss_{pred,suiv}}{\partial y} > 0$, $y'$ tend vers une valeur plus petite. Il sera donc situé plus haut dans l'image. Des conclusions équivalentes sont obtenues en regardant le gradient $\frac{\partial Loss_{pred,pré}}{\partial y}$.

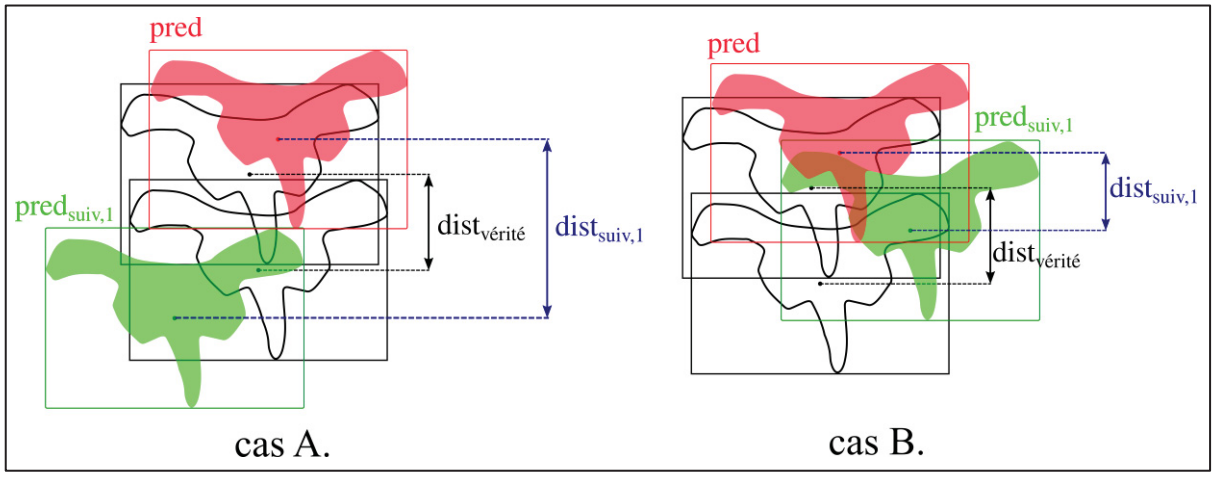



Figure 4.2 Cas d'application de pénalités de distances intervertébrales

### 4.2.3 Méthode d'évaluation des stratégies de spécialisation

Afin de connaître l'intérêt des stratégies de spécialisation, chacune d'entre elles est intégrée individuellement à DetectoRS. DetectoRS est alors entrainé et validé sur l'ensemble d'entrainement selon la stratégie de validation croisée à cinq plis, de la même manière que dans le CHAPITRE 3.

Les résultats en validation croisée permettent de catégoriser la stratégie comme bénéfique, inutile ou néfaste.

Les stratégies de spécialisation jugées comme bénéfiques sont alors intégrées à DetectoRS pour former le CNN V-DetectoRS. Les versions de DetectoRS intégrant les stratégies bénéfiques prises individuellement et V-DetectoRS sont évalués sur l'ensemble de test. Pour connaître l'influence de chacune des stratégies d'amélioration, l'analyse des résultats des différentes versions de DetectoRS sur l'ensemble de test est effectuée.

## 4.3 Résultats

### 4.3.1 Résultats en validation croisée

#### 4.3.1.1 Résultats du pré-entrainement ajusté en validation croisée

L'entrainement de ResNet50 pour la tâche de classification des images ChestX-ray14 est fait sur six époques. ResNet50 atteint alors une valeur d'AUC (*Area Under the ROC Curve*) en test de 81,6%.

Les résultats en validation croisée des trois configurations de pré-entrainement, soit DetectoRS pré-entrainé sur ImageNet, DetectoRS pré-entrainé sur ChestX-ray14 avec les poids des couches basses fixés et DetectoRS pré-entrainé sur ChestX-ray14 avec les poids comme initialisation, sont donnés dans le Tableau 4.1

Tableau 4.1 Comparaison des résultats obtenus en validation avec le pré-entrainement sur ImageNet et les pré-entrainements sur ChestX-ray14

| | Métriques d'évaluation (moyenne ± écart-type sur les 5plis) | | | | |
|---|---|---|---|---|---|
| Configuration | Dice | $IR_{D\geq 50}$ | $IR_{D\geq 75}$ | $IR_{d\leq 10}$ | $IR_{d\leq 5}$ |
| Pré-entrainement ImageNet | 85,89±0,94 | 95,58±1,45 | 94,00±1,52 | 95,21±1,28 | 90,81±1,17 |
| Pré-entrainement ChestX-ray14 - Poids des couches basses fixés | 85,21±1,09 | 95,19±1,62 | 93,31±1,71 | 94,69±1,51 | 89,11±1,49 |
| Pré-entrainement ChestX-ray14 - Poids comme initialisation | 85,36±1,45 | 94,95±1,65 | 93,41±1,83 | 94,59±1,59 | 89,25±1,59 |

Les performances des deux configurations de pré-entrainement sur ChestX-ray14 sont très similaires et sont légèrement en dessous de celles atteintes par DetectoRS en pré-entrainement sur ImageNet. De plus, le pré-entrainement sur ChestX-ray14 requiert un temps d'entrainement plus long.

#### 4.3.1.2 Résultats du post-traitement anatomiquement contraignant en validation croisée

Les résultats après application du post-traitement initial et du post-traitement anatomiquement contraignant en validation croisée sont donnés dans le Tableau 4.2.

Tableau 4.2 Comparaison des résultats obtenus sur les deux types de post-traitement en validation croisée

| | Moyennes sur les métriques d'évaluation (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Dice | $IR_{D\geq 50}$ | $SR_{D\geq 50}$ | $IR_{D\geq 75}$ | $SR_{D\geq 75}$ | $IR_{d\leq 10}$ | $SR_{d\leq 10}$ | $IR_{d\leq 5}$ | $SR_{d\leq 5}$ |
| Post-traitement naïf | 85,89 | 95,58 | 84,48 | 94,00 | 77,11 | 95,21 | 83,56 | 90,81 | 56,68 |
| Post-traitement anatomiquement contraignant | 86,11 | 95,91 | 92,17 | 94,30 | 82,03 | 95,66 | 91,40 | 91,23 | 59,60 |
| Gain | 0,26 | 0,35 | 9,10 | 0,32 | 6,38 | 0,47 | 9,38 | 0,46 | 5,15 |

Le post-traitement anatomiquement contraignant offre de meilleurs résultats sur toutes les métriques par rapport au post-traitement naïf. Le post-traitement intervertébral permet de se rapprocher des performances obtenues quand les véritables meilleures prédictions sont sélectionnées, même s'il reste tout de même un écart. Sur les résultats en validation croisée, le choix de 414 prédictions a été révisé, correspondant à la rectification de 108 images sur 651. Les gains sur les métriques de segmentation et d'identification ne sont pas particulièrement élevés. Cependant, les écarts de performance sur les métriques de taux de succès sont beaucoup plus marqués : le gain minimal est de 5,15% pour $SR_{d\leq5}$ et le gain maximal est de 9,38% pour $SR_{d\leq10}$.

Les résultats relatifs à l'efficacité du post-traitement anatomiquement contraignant sont donnés dans le Tableau 4.3. L'efficacité varie entre 88% pour $SR_{D\geq50}$ et 100% pour $SR_{d\leq5}$.

Tableau 4.3 Efficacité du post-traitement anatomiquement contraignant sur les résultats en validation croisée

| | $SR_{D\geq50}$ | $SR_{D\geq75}$ | $SR_{d\leq10}$ | $SR_{d\leq5}$ |
|---|---|---|---|---|
| Nombre d'images dégradées | 8,00 | 3,00 | 4,00 | 0,00 |
| Nombre d'images améliorées | 58,00 | 35,00 | 55,00 | 19,00 |
| Efficacité (%) | 87,88 | 92 | 93 | 100 |

Un exemple de correction d'une image qui présentait des anomalies avec le post-traitement naïf est donné en Figure 4.3. Avant le post-traitement anatomiquement contraignant, l'image montre T7 et T8 superposées et T11 manquée. Le post-traitement, en rectifiant le choix de T8, T9, T10 et T11, produit un résultat anatomiquement plausible. L'ANNEXE III montre deux autres exemples de correction.

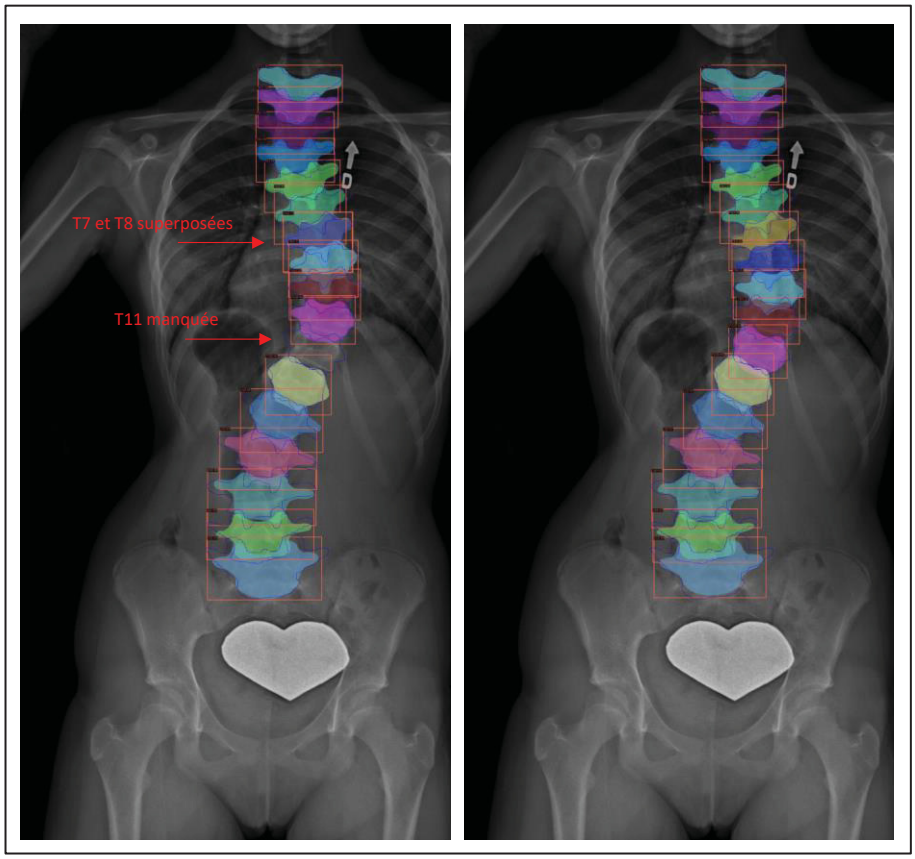



Figure 4.3 Exemple de correction d'une image avant (à gauche) et après (à droite) le post-traitement anatomiquement contraignant

#### 4.3.1.3 Résultats de la fonction de perte anatomiquement contraignante en validation croisée

Pour connaître l'influence du terme de pénalité, le terme initial de régression de la fonction de perte ainsi que le terme de pénalité ajouté sont calculés au cours de l'entrainement et de la validation. La Figure 4.4 affiche les deux termes de perte en entrainement et en validation sur un plis. L'allure de la courbe du terme de pénalité en entrainement converge bien vers une valeur minimale, mais sa pente est moins forte que celle de la courbe du terme de régression. On observe aussi du bruit sur la courbe du terme de pénalité. L'entrainement est arrêté à 5000 itérations. DetectoRS est entrainé et validé sur la nouvelle fonction de perte en régression, intégrant le terme $Loss_{InterV}$, en respectant la même stratégie de validation croisée en cinq

plis, établie au CHAPITRE 3. Différents entrainements sont effectués selon différentes valeurs de $\lambda$. La valeur de $\lambda$ optimale est trouvée pour $\lambda = 0{,}1$.

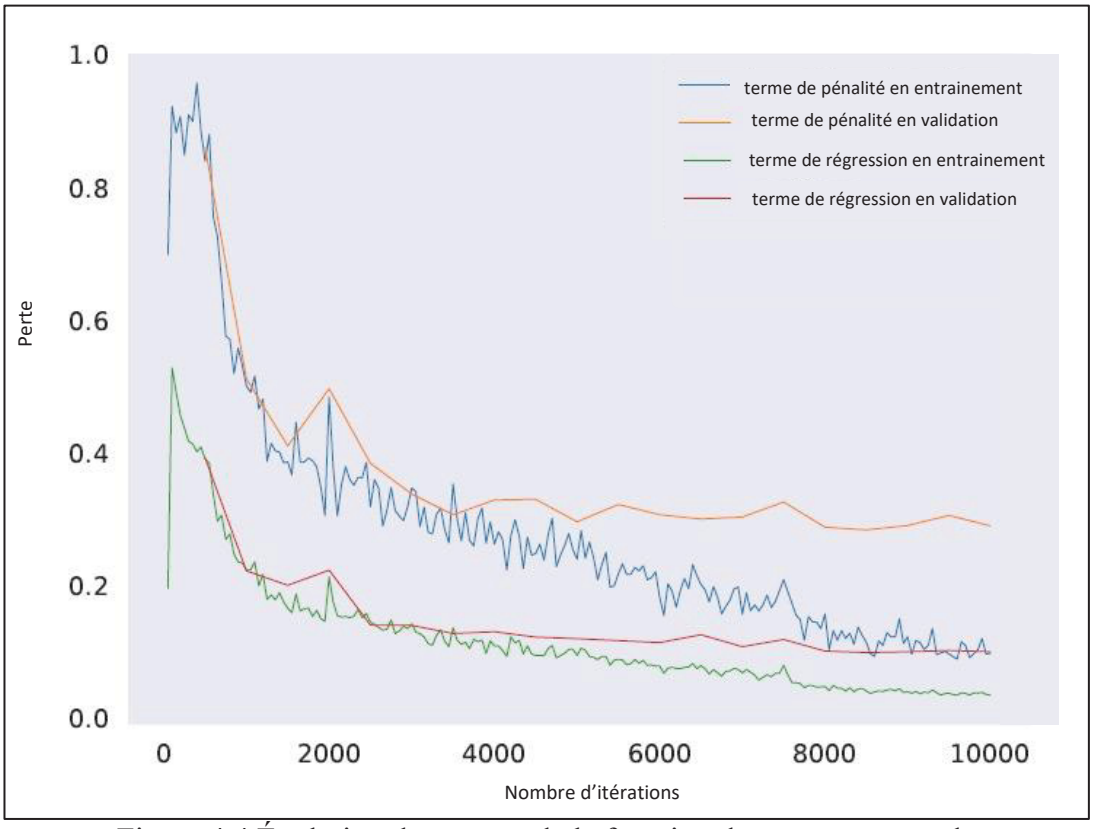



Figure 4.4 Évolution des termes de la fonction de perte au cours de l'entrainement et de la validation

Les résultats moyens sur les métriques d'évaluation sont donnés dans le Tableau 4.4. L'intégration du terme de pénalité $Loss_{IntervV}$ offre des gains de performances sur toutes les métriques d'évaluation sauf pour $IR_{d\leq5}$ et $SR_{d\leq5}$. Sur ces taux d'identification et de succès relatifs à la condition d'identification la plus contraignante, la modification de la fonction de perte dégrade légèrement les résultats. Les gains de performances sont plus marqués sur les taux de succès $SR_{d\leq10}$, $SR_{D\geq50}$ et $SR_{D\geq75}$.

Tableau 4.4 Comparaison des performances obtenues avec la fonction de perte standard et celle intégrant le terme de pénalité en validation

| | Moyennes sur les métriques d'évaluation (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Dice | $IR_{D\geq 50}$ | $SR_{D\geq 50}$ | $IR_{D\geq 75}$ | $SR_{D\geq 75}$ | $IR_{d\leq 10}$ | $SR_{d\leq 10}$ | $IR_{d\leq 5}$ | $SR_{d\leq 5}$ |
| Fonction de perte standard | 85,89 | 95,58 | 84,48 | 94,00 | 77,11 | 95,21 | 83,56 | 90,81 | 56,68 |
| Fonction de perte avec terme de pénalité | 86,29 | 95,59 | 86,08 | 94,12 | 79,25 | 95,29 | 85,25 | 90,60 | 55,60 |
| Gain | 0,47 | 0,01 | 1,89 | 0,13 | 2,78 | 0,08 | 2,02 | -0,23 | -1,91 |

### 4.3.2 Évaluation de V-DetectoRS sur l'ensemble de test

Les propositions d'amélioration ne sont pas toutes intégrées à DetectoRS. En effet, le pré-entrainement ajusté ne s'est pas montré bénéfique sur l'ensemble de validation. En revanche, le post-traitement ainsi que la fonction de perte anatomiquement contraignante permettent des gains de performances. Le CNN DetectoRS intégrant les deux stratégies d'amélioration est nommé V-DetectoRS.

Tableau 4.5 Résultats des quatre configurations de DetectoRS sur l'ensemble de test

| | Moyennes sur les métriques d'évaluation (%) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Dice | $IR_{D\geq 50}$ | $SR_{D\geq 50}$ | $IR_{D\geq 75}$ | $SR_{D\geq 75}$ | $IR_{d\leq 10}$ | $SR_{d\leq 10}$ | $IR_{d\leq 5}$ | $SR_{d\leq 5}$ |
| DetectoRS | 86,36 | 95,79 | 85,34 | 94,88 | 80,17 | 95,54 | 83,62 | 92,09 | 55,17 |
| DetectoRS avec terme de pénalité intervertébrale | 86,26 | 94,98 | 82,76 | 94,12 | 78,45 | 94,88 | 82,76 | 91,99 | 60,34 |
| DetectoRS avec post-traitement intervertébral | 86,31 | 95,84 | 93,97 | 94,68 | 84,48 | 95,64 | 92,24 | 91,19 | 58,62 |
| V-DetectoRS | 87,84 | 97,46 | 95,69 | 96,35 | 86,21 | 97,36 | 94,83 | 93,76 | 66,38 |

Les résultats de quatre configurations de DetectoRS sur l'ensemble de test sont rassemblés dans le Tableau 4.5 : V-DetectoRS est évalué sur l'ensemble de test afin d'évaluer ses capacités de généralisation sur un ensemble de données inconnu; les deux propositions d'amélioration

sont aussi évaluées individuellement sur l'ensemble de test afin de connaître l'influence de chacune et les résultats de DetectoRS sur l'ensemble de test, calculés au CHAPITRE 3, sont rappelés afin de pouvoir être comparés.

Les distributions des images résultantes selon les métriques de segmentation et d'identification sont données pour les quatre configurations en Figure 4.5.

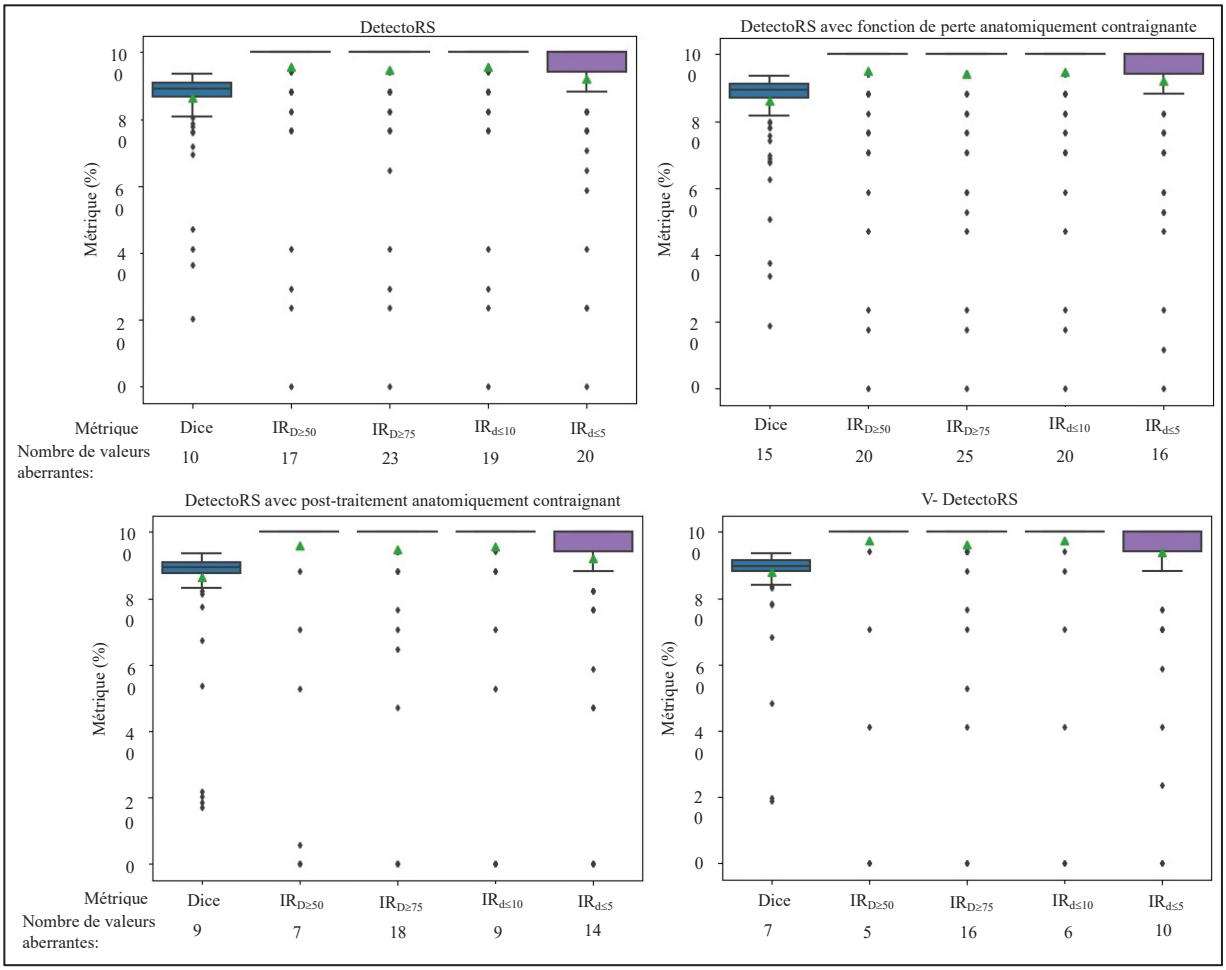



Figure 4.5 Distribution des images résultantes pour les métriques de segmentation et d'identification pour les quatre versions de DetectoRS évaluées sur l'ensemble de test

Prises séparément, les améliorations proposées ne performent pas aussi bien sur l'ensemble de test qu'en validation croisée à cinq plis.

DetectoRS avec le terme de pénalité intervertébrale présente même des résultats légèrement en baisse par rapport au DetectoRS standard sur toutes les métriques sauf $SR_{d\leq 5}$. Sur cette dernière, une amélioration de 5,17% est observée.

Au regard de la Figure 4.5, la distribution des images résultantes de DetectoRS avec la fonction de perte anatomiquement contraignante est assez similaire à celle de DetectoRS standard. La principale différence tient dans le fait que DetectoRS intégrant la fonction de perte ajustée produit plus d'images avec des identifications erronées, identifiables par les valeurs aberrantes, que DetectoRS standard.

Cependant, les performances maximales potentielles des taux de succès, calculées en sélectionnant les véritables meilleures prédictions, sont légèrement plus élevées pour DetectoRS pris avec le terme de pénalité que pour DetectoRS standard. Elles sont données dans le Tableau 4.6.

Tableau 4.6 Performances de taux de succès potentiels sur l'ensemble de test

| | Moyennes sur les taux de succès potentiels (%) | | | |
|---|---|---|---|---|
| | $SR_{D\geq 50}$ | $SR_{D\geq 75}$ | $SR_{d\leq 10}$ | $SR_{d\leq 5}$ |
| DetectoRS | 94,83 | 85,34 | 93,10 | 59,49 |
| DetectoRS avec terme de pénalité intervertébral | 96,55 | 86,21 | 94,83 | 66,38 |

Le post-traitement anatomiquement contraignant n'améliore pas les performances sur les métriques de segmentation et d'identification des vertèbres (Tableau 4.5). Elles sont même un peu en retrait pour Dice, $IR_{d\leq 5}$ et $IR_{D\geq 75}$. Sur les métriques de taux de succès, DetectoRS pris avec le post-traitement intervertébral offre une hausse de résultats comprise entre 5,89% et 10,31%. Les distributions des images résultantes sur les métriques d'identification, qui présentent moins de valeurs aberrantes, témoignent aussi de la correction d'identification dans les images.

Le post-traitement a rectifié 64 vertèbres, correspondant à la modification de 19 images sur les 116 images de l'ensemble de test. Les nombres d'images dégradées (passage du statut « succès » à « échec ») et d'images améliorées (passage du statut « échec » à « succès ») ainsi

que l'efficacité (nombre d'images améliorées sur le nombre total d'images modifiées) sont donnés pour chaque métrique de taux de succès dans le Tableau 4.7. L'efficacité varie selon les métriques entre 81% et 86%.

Tableau 4.7 Efficacité du post-traitement anatomiquement contraignant pour DetectoRS en test

| | $SR_{D\geq 50}$ | $SR_{D\geq 75}$ | $SR_{d\leq 10}$ | $SR_{d\leq 5}$ |
|---|---|---|---|---|
| Nombre d'images dégradées | 3,00 | 1,00 | 2,00 | 1,00 |
| Nombre d'images améliorées | 13,00 | 6,00 | 12,00 | 5,00 |
| Efficacité (%) | 81,25 | 85,71 | 85,71 | 83,33 |

V-Detectors présente systématiquement les meilleures performances par rapport aux autres configurations (Tableau 4.5). Le coefficient de Dice perçoit un gain de 1,71% par rapport à DetectoRS standard. Sur les métriques d'identification des vertèbres, l'amélioration est comprise entre 1,55% et 1,90%. Enfin, pour les métriques de taux de succès, le gain est compris entre 7,53% pour $SR_{D\geq 75}$ et 20,32% pour $SR_{d\leq 5}$.

Les taux de succès en hausse se traduisent par la réduction du nombre de valeurs aberrantes dans les graphes de distributions des images résultantes selon les métriques d'identification (Figure 4.5).

Sur les 116 images, 20 ont été modifiées par le post-traitement, correspondant à la révision de 70 prédictions de vertèbres. Le Tableau 4.8 montre des scores d'efficacité de V-Detectors. Par rapport à DetectoRS intégrant seulement le post-traitement anatomiquement contraignant, les scores sont en hausse pour les quatre métriques.

Tableau 4.8 Efficacité du post-traitement anatomiquement contraignant pour V-DetectoRS en test

| | $SR_{D\geq 50}$ | $SR_{D\geq 75}$ | $SR_{d\leq 10}$ | $SR_{d\leq 5}$ |
|---|---|---|---|---|
| Nombre d'images dégradées | 1,00 | 0,00 | 1,00 | 0,00 |
| Nombre d'images améliorées | 16,00 | 9,00 | 15,00 | 7,00 |
| Efficacité (%) | 94,12 | 100 | 93,75 | 100 |

## 4.4 Discussion

### 4.4.1 Analyse du pré-entrainement ajusté

Les résultats en validation croisée (Tableau 4.1) montrent que le pré-entrainement sur ChestX-ray14 est moins efficace que le pré-entrainement sur ImageNet : il ne permet pas d'amélioration des résultats et impose un temps d'entrainement plus long. Une potentielle raison de l'inefficacité du pré-entrainement sur ChestX-ray14 tient dans la nature de la tâche effectuée en pré-entrainement. Sur ChestX-ray14, ResNet50 effectue une simple classification. ResNet50 n'est donc pas entrainé pour localiser et segmenter des objets dans des images. Au contraire, le pré-entrainement de ResNet50 sur ImageNet lui permet d'apprendre les tâches de localisation et de segmentation sur des images optiques. Le pré-entrainement sur ChestX-ray14 correspond au bon domaine de données, mais à la mauvaise tâche.

### 4.4.2 Analyse du post-traitement anatomiquement contraignant

Les résultats en validation croisée montrent qu'avec le post-traitement anatomiquement contraignant, le choix des prédictions est mieux réalisé qu'avec le post-traitement initial (Tableau 3.8). Les taux de succès en hausse témoignent de la rectification d'images qui présentaient des anomalies. Les contraintes de respect des distances intervertébrales permettent de corriger quelques prédictions mal sélectionnées, qui font passer des images du statut « en échec » au statut « en succès ». Le post-traitement est capable de corriger des anomalies en cascade ainsi que des anomalies sur les vertèbres situées aux extrémités.

L'augmentation notable des taux de succès offert par le post-traitement anatomiquement contraignant par rapport au post-traitement naïf sur l'ensemble de test signifie que les intervalles de tolérance établis à partir des données d'entrainement sont relativement bien généralisables à d'autres données. Cependant, le post-traitement anatomiquement contraignant en sortie de DetectoRS est moins efficace sur l'ensemble de test que sur l'ensemble d'entrainement (Tableau 4.3 et Tableau 4.7). L'objectif du post-traitement est tout de même rempli : la proportion d'images résultantes utilisables a été augmentée.

### 4.4.3 Analyse de la fonction de perte anatomiquement contraignante

En validation croisée, la fonction de perte anatomiquement contraignante permet une augmentation des taux de succès (Tableau 4.4). Ce résultat est conforme à l'objectif du terme de pénalité : les distances intervertébrales étant mieux respectées, il y a moins d'anomalies sur les images, entrainant un taux de succès en hausse.

Par comparaison des résultats, on observe que le comportement de DetectoRS avec le terme de pénalité en validation croisée est différent de DetectoRS avec le terme de pénalité en test (Tableau 4.4 et Tableau 4.5). Bien que le terme de pénalité semble bénéfique en validation croisée, les résultats en test ne semblent pas satisfaisants. Le terme de pénalité ne semble pas bien généraliser aux images de l'ensemble de test. Il produit des images résultantes anormales supplémentaires, identifiables comme valeurs aberrantes sur la Figure 4.5 et qui sont responsables des résultats moyens détériorés. Cependant, la fonction de perte permet de prédire des résultats très précis en localisation, comme le témoigne la valeur à la hausse de la métrique $SR_{d\leq5}$.

Par rapport à DetectoRS standard, DetectoRS intégrant le terme de pénalité intervertébral permet de produire de meilleurs résultats potentiels (Tableau 4.6). Cela signifie que DetectoRS intégrant la fonction de perte anatomiquement contraignante produit des prédictions bien localisées et identifiées, mais qui ne présentent pas le score de confiance le plus élevé. Pour DetectoRS intégrant le terme de pénalité, 4,11% des prédictions qui présentent le plus haut score de confiance ne sont en vérité pas les meilleures prédictions contre 3,8% pour DetectoRS standard.

Les erreurs de classification, plus spécifiquement d'attribution du score de confiance, sont responsables de résultats pratiques dégradés. L'ajout du terme de pénalité dans la fonction de perte permet de produire des prédictions plus précises en localisation, mais dégrade légèrement les performances de classification, plus spécifiquement de l'attribution du score de confiance. À cause de ces erreurs de classification, DetectoRS intégrant une fonction de perte basée sur

des contraintes de distances intervertébrales ne permet pas de produire des résultats plus anatomiquement plausibles que DetectoRS standard sur l'ensemble de test.
Plusieurs hypothèses peuvent être faites quant aux causes des erreurs de classification faites par DetectoRS intégrant la fonction de perte sur les données test. La définition complexe du terme de pénalité le rend plus difficile à apprendre et peu gêner l'apprentissage relatif à la fonction de perte de classification. En général, les fonctions de pertes simples et continues sont préférées aux fonctions complexes. De plus, la fonction de perte intervertébrale mise en place a tendance à pénaliser les prédictions correctes si leurs voisines sont incorrectes, perturbant encore l'apprentissage.

### 4.4.4 Analyse de V-DetectoRS

L'utilisation conjointe de la fonction de perte et du post-traitement anatomiquement contraignants semble très bénéfique : V-DetectoRS domine sur toutes les métriques d'évaluation (Tableau 4.5). Il a été vu que DetectoRS pris avec le terme de pénalité produit des prédictions de bonne qualité, mais attribue mal leur score de confiance : la tâche de classification n'est pas très bien réalisée. Le post-traitement permet de rectifier cette erreur en s'appuyant sur le respect des distances intervertébrales. Il est particulièrement efficace sur les résultats en sortie de DetectoRS intégrant le terme de pénalité intervertébrale. Ce gain d'efficacité peut en partie s'expliquer par le fait que, quand la prédiction est mal choisie, la réelle meilleure prédiction est celle avec le deuxième plus haut score de confiance dans 96,3% des cas pour DetectoRS intégrant le terme de pénalité, contre seulement 85,3% des cas pour DetectoRS standard.

## 4.5 Conclusion

Le travail effectué dans ce chapitre est consacré à la spécialisation de DetectoRS à la tâche de segmentation et d'identification des vertèbres sur des radiographies frontales EOS. Cet ajustement est réalisé dans le but d'obtenir de meilleures performances. Une analyse des

erreurs de DetectoRS plus poussée est conduite afin de mettre en place des stratégies d'amélioration pertinentes. L'analyse conduite montre notamment qu'il existe des images résultantes montrant des anomalies : des vertèbres sont manquées et des prédictions sont superposées. Ces images anormales présentent des taux d'identification très bas et font chuter les résultats d'évaluation moyens. Elles sont aussi en partie responsables de faibles taux de succès.

Trois stratégies d'amélioration sont proposées. Elles reposent toutes sur l'intégration de connaissances préalables de contexte supplémentaires. La première proposition consiste en un pré-entrainement ajusté aux images radiographiques. La deuxième stratégie d'amélioration proposée est un post-traitement anatomiquement contraignant et la dernière stratégie d'amélioration consiste en l'intégration des contraintes anatomiques de distances intervertébrales au sein de l'apprentissage.

Après l'évaluation des trois stratégies en validation croisée, le post-traitement et la fonction de perte modifiée sont intégrés à DetectoRS, formant le CNN ajusté à la tâche V-DetectoRS. V-DetectoRS est évalué sur l'ensemble de test afin de connaître ses capacités de généralisation. L'intégration des deux stratégies d'amélioration conjointes permet une hausse notable des résultats de V-DetectoRS par rapport à DetectoRS étudié dans le CHAPITRE 3. Le coefficient de Dice atteint un score de 87,84% et, en considérant les vertèbres dont le coefficient de Dice est supérieur à 75% comme bien identifiées, 86,21% des résultats sont directement utilisables en contexte clinique.

# DISCUSSION GÉNÉRALE

L'objectif principal de ce travail était de proposer une méthode automatique de segmentation et d'identification individuelle des vertèbres thoraciques et lombaires sur des radiographies frontales EOS de patient atteint de SIA. Ce travail s'inscrit dans le projet d'automatisation des applications de l'entreprise EOS Imaging.

L'étude se tourne vers les méthodes d'apprentissage profond. Quatre CNNs de segmentation d'instance, DetectoRS, MS R-CNN, RetinaMask et YOLACT, sont comparés sur la tâche d'identification et de segmentation des vertèbres sur les radiographies frontales EOS. Les CNNs sont évalués sur des métriques médicales de Dice, qualifiant la qualité de la segmentation, et de taux d'identification. Plusieurs taux d'identification sont introduits. Ils correspondent à différentes conditions d'identification, plus ou moins contraignantes. Un post-traitement est mis en place pour conserver une unique prédiction par vertèbre à détecter. L'étude comparative montre que DetectoRS est légèrement meilleur que les autres CNNs. Il se démarque surtout sur les taux d'identification avec des conditions particulièrement contraignantes. MS R-CNN et RetinaMask performent de manière similaire tandis que YOLACT se révèle nettement moins bon que les trois autres CNNs.

Dans une deuxième partie, on cherche à améliorer les performances de DetectoRS en intégrant des connaissances préalables à la méthode. L'intérêt d'ajouter des connaissances préalables est souligné dans la revue de littérature. Les méthodes CNN spécifiquement conçues pour la segmentation et l'identification de vertèbres sur des images radiologiques intègrent toutes des informations relatives à la structure anatomique du rachis (Lessmann et al., 2019 ; Payer et al., 2019 ; Sekuboyina et al., 2021 ; Kim et al., 2020). Trois stratégies d'amélioration sont mises en place : un pré-entrainement ajusté, un post-traitement anatomiquement contraignant et l'ajout d'un terme de pénalité dans la fonction de perte d'entrainement de DetectoRS. Ces stratégies sont testées sur une validation croisée à 5 plis, comme dans le CHAPITRE 3.

Le pré-entrainement, qui utilise des radiographies du thorax en tant qu'images sources, ne se révèle pas efficace. La différence de tâche apprise entre le pré-entrainement et l'entrainement est en certainement la cause.

Le post-traitement tenant compte des distances intervertébrales remplace le post-traitement naïf utilisé dans l'étude comparative. Le post-traitement, en assurant au maximum le respect des distances intervertébrales, sélectionne mieux les prédictions à conserver. Il permet une large augmentation des performances sur les métriques de taux de succès, spécifiquement quand la condition d'identification est souple.

L'entrainement de DetectoRS est aussi soumis à des contraintes de distances intervertébrales. Elles sont ajoutées sous la forme d'un terme de pénalité dans la fonction de perte de régression des boîtes englobantes. Les pénalités sont appliquées aux prédictions qui ne respectent pas des distances intervertébrales vraisemblables avec les prédictions adjacentes. Cette fonction de perte ajustée assure une très légère amélioration des performances, notamment au niveau de la proportion d'images résultantes dont toutes les vertèbres sont bien détectées. L'analyse des pertes en entrainement et validation montre que l'entrainement est effectué, mais est soumis à des instabilités. Cette instabilité est certainement due à la complexité de définition du terme de pénalité.

Le post-traitement et la fonction de perte anatomiquement contraignants, propositions d'amélioration jugées comme bénéfiques, sont intégrées à DetectoRS. Cette version de DetectoRS ajustée à la tâche de segmentation et d'identification des vertèbres sur les radiographies frontales EOS est nommée V-DetectoRS. V-DetectoRS ainsi que DetectoRS avec les deux stratégies d'amélioration prises individuellement sont évalués sur l'ensemble de test afin de connaître leur réel intérêt.

DetectoRS intégrant la fonction de perte anatomiquement contraignante performe moins bien que DetectoRS. En revanche, le post-traitement anatomiquement contraignant permet un gain de performances par rapport à DetectoRS avec le post-traitement initial. V-DetectoRS, qui joint les deux stratégies d'amélioration, performe nettement mieux que les trois autres configurations de DetectoRS. La fonction de perte permet de générer des prédictions de bonne qualité, mais dont les scores de confiance sont un peu moins fiables. La tâche de classification est donc moins bien réalisée que pour DetectoRS. Le post-traitement anatomiquement contraignant rectifie cette erreur en sélectionnant les prédictions qui respectent au mieux les contraintes de distances intervertébrales.

En considérant qu'une vertèbre est bien identifiée lorsqu'elle montre un coefficient de Dice de 75%, 86,21% des images testées ont été traitées par V-DetectoRS avec succès.

Par rapport aux études discutées dans la revue de littérature, ce travail présente les particularités majeures d'être effectué sur des radiographies en vue frontale et de considérer les vertèbres entières, non pas seulement le corps vertébral. Ces conditions rendent les tâches de segmentation et d'identification plus complexes. D'abord les objets à détecter et segmenter sont particulièrement superposés : le taux de superposition moyen est de 15,6%. De plus, la prise en compte des apophyses dans les masques de segmentation donne une forme plus complexe à ces derniers. On note aussi que la visibilité des apophyses est particulièrement restreinte sur les radiographies frontales. Bien qu'il ne soit pas possible de faire une comparaison directe des performances, on propose de mettre en perspective les résultats obtenus.

Le coefficient de Dice de 87,84% atteint par V-DetectoRS sur l'ensemble de test est similaire à ceux obtenus par Mask R-CNN dans l'étude de (Yang et al., 2019), soit 89,8% et Mask R-CNN et YOLACT dans l'étude de (Kónya et al., 2021), soit respectivement 85,87% et 87,77%. La méthode de (Kim *et al.*, 2020), qui traite aussi de images radiographiques, obtient un meilleur coefficient de Dice pour la segmentation des lombaires. Cet écart de 3,76 pp peut s'expliquer par le fait que, dans l'étude de (Kim *et al.*, 2021), le champ de vision intègre seulement les cinq lombaires et les apophyses ne sont pas présentes dans les masques de segmentation.

Par rapport aux coefficients de Dice moyens des méthodes VerSe appliquées au données CT-scans (Lessmann *et al.*, 2019 ; Payer *et al.*, 2019 ; Sekuboyina *et al.*, 2021) donnés dans le Tableau 1.1, le coefficient de Dice de V-DetectoRS est aussi en retrait. Ces différences s'expliquent en partie par les différences entre les données d'entrées : les CT-scans sont des images 3D qui comportent beaucoup plus d'informations de contexte et pas de superposition des structures. En revanche, le taux d'identification selon une contrainte de 10 mm de V-DetectoRS est plus élevé que les taux d'identification selon une contrainte de 20 mm des méthodes VerSe. Cet écart est très probablement une conséquence de l'identification individuelle de toutes les vertèbres du rachis (et non pas seulement des thoraciques et

lombaires) et de la présence de cas sur- et sous-numéraires dans le concours VerSe, qui complexifient la tâche d'identification.

L'étude proposée par (Yang et al., 2019) est relativement proche de celle effectuée dans ce travail : des radiographies EOS de patients scoliotiques, notamment en vue frontale, sont traitées avec un CNN de segmentation d'instance populaire (Mask R-CNN). La majeure différence tient dans l'utilisation combinée des radiographies selon les vues frontales et latérales, qui apporte des informations complémentaires par rapport à l'utilisation d'une seule vue. Le post-traitement mis en place dans ce travail, sélectionnant une unique prédiction par vertèbre à détecter, permet d'adopter la métrique du taux d'identification plutôt que les métriques de précision et de justesse utilisées par (Yang et al., 2019). Le taux d'identification avec condition d'un coefficient de Dice supérieur ou égal à 50% atteint 97,46%. Il est similaire aux scores de précision et justesse obtenus par (Yang et al., 2019), indiqués dans le Tableau 1.2. Cependant, contrairement à (Yang et al., 2019), les images radiographiques ne sont pas recadrées autour du rachis à l'aide des données d'annotations. Les conditions d'utilisation du CNN en pratique sont donc mieux respectées. Malgré la difficulté apportée par la prise en compte des apophyses et l'identification individuelle des vertèbres, les coefficients de Dice atteints sont similaires.

# CONCLUSION ET RECOMMANDATIONS

## 6.1 Conclusions

Ce mémoire a permis de proposer une méthode automatique de segmentation et d'identification individuelle des vertèbres sur des radiographies frontales EOS de rachis. La méthode finale V-DetectoRS est basée sur le CNN de segmentation d'instance DetectoRS. Elle intègre une fonction de perte en régression et un post-traitement anatomiquement contraignants qui modélisent les distances intervertébrales. V-DetectoRS atteint sur l'ensemble de test un coefficient de Dice moyen de 87,84%. En considérant qu'en pratique, une vertèbre est bien identifiée lorsqu'elle montre un coefficient de Dice de 75%, 86,21% des images testées ont été traitées avec succès par V-DetectoRS.

Ce travail, effectué dans le contexte d'automatisation d'applications de l'entreprise EOS Imaging, pourrait dans le futur être intégré aux processus de détermination de métriques cliniques et de reconstruction 3D de rachis depuis des radiographies EOS. Plus spécifiquement, dans le processus de reconstruction 3D du rachis, les informations d'identification et de segmentation des vertèbres permettent d'initialiser la position des corps vertébraux et d'affiner la forme des modèles 3D des vertèbres.

## 6.2 Recommandations

Quelques pistes d'amélioration de V-DetectoRS pourraient être envisagées. Évidemment, travailler sur un ensemble de données plus large permettrait des performances en hausse. Plus spécifiquement, une meilleure représentation des cas de scolioses plus rares permettrait de mieux traiter ces derniers. De plus, les intervalles de tolérances relatifs aux distances intervertébrales pourraient être ajustés en étant calculés sur de plus grands échantillons. Au niveau de la modification de la fonction de perte, il pourrait être envisageable d'intégrer un terme de pénalité dans la fonction relative à la classification des vertèbres. En effet, V-

DetectoRS produit des erreurs de classification : les prédictions sont spatialement précises, mais leurs scores de confiances sont moins fiables qu'avec DetectoRS. L'intégration d'un terme relatif aux distances dans la fonction de perte de classification semble tout de même complexe.

Une autre possibilité serait d'utiliser une machine à vecteurs de support (SVM pour Support Vector Machine) en sortie de DetectoRS pour effectuer la tâche d'identification des vertèbres à partir des prédictions. La difficulté tiendrait dans la sélection des 17 vertèbres parmi toutes celles prédites.

# ANNEXE I

## RÉSULTATS EN VALIDATION CROISÉE POUR LES QUATRE CNNS

Tableau-A I-1

| | DetectoRS | | | | | |
|---|---|---|---|---|---|---|
| | Dice | $IR_{D\geq 50}$ | $IR_{D\geq 75}$ | $IR_{d\leq 10}$ | $IR_{d\leq 5}$ | Score |
| Config 1: η = 0,0025 ; B = 2 | 85,02±0,80 | 95,12±1,25 | 93,37±1,06 | 94,76±1,10 | 89,72±0,86 | 89,13 |
| Config 2: η = 0,0025 ; B = 1 | 85,33±0,91 | 95,73±1,45 | 94,13±1,42 | 95,33±1,37 | 90,06±1,43 | 89,57 |
| Config 3: η = 0,0075 ; B = 1 | DIVERGE SUR UN PLIS | | | | | 0,00 |
| Config 4: η = 0,0050 ; B = 1 | 85,89±1,13 | 95,58±1,67 | 94,00±1,63 | 95,21±1,69 | 90,81±1,65 | 89,89 |
| | MS R-CNN | | | | | |
| | Dice | $IR_{D\geq 50}$ | $IR_{D\geq 75}$ | $IR_{d\leq 10}$ | $IR_{d\leq 5}$ | Score |
| Config 1: η = 0,0025 ; B = 2 | 84,39±1,21 | 95,22±1,83 | 90,55±2,48 | 93,57±2,31 | 82,77±2,25 | 87,46 |
| Config 2: η = 0,0025 ; B = 1 | 84,05±1,08 | 95,16±1,55 | 90,92±1,73 | 93,53±1,73 | 81,89±2,30 | 87,21 |
| Config 3: η = 0,0050 ; B = 1 | DIVERGE SUR UN PLIS | | | | | 0,00 |
| Config 4: η = 0,0050 ; B = 2 | 85,13±0,86 | 95,37±1,43 | 92,07±1,46 | 94,07±1,43 | 85,16±2,04 | 88,40 |
| Config 5: η = 0,0075 ; B = 2 | DIVERGE SUR UN PLIS | | | | | 0,00 |
| | RetinaMask | | | | | |
| | Dice | $IR_{D\geq 50}$ | $IR_{D\geq 75}$ | $IR_{d\leq 10}$ | $IR_{d\leq 5}$ | Score |
| Config 1: η = 0,0025 ; B = 2 | 84,90±0,95 | 94,58±1,45 | 91,69±1,82 | 93,86±1,44 | 86,23±1,89 | 88,25 |
| Config 2: η = 0,0025 ; B = 1 | 85,01±0,97 | 94,82±1,46 | 92,34±1,87 | 94,10±1,64 | 85,87±2,44 | 88,40 |
| Config 3: η = 0,0050 ; B = 1 | 85,51±1,16 | 95,19±1,66 | 92,68±1,80 | 94,38±1,75 | 86,98±2,38 | 88,91 |
| Config 4: η = 0,0075 ; B = 1 | 85,57±1,08 | 94,92±1,63 | 92,81±1,70 | 94,32±1,67 | 87,68±2,15 | 89,00 |
| Config 5 : η = 0,01 ; B = 1 | 85,57±1,12 | 94,83±1,64 | 92,89±2,01 | 94,14±1,75 | 87,28±2,55 | 88,93 |
| | YOLACT | | | | | |
| | Dice | $IR_{D\geq 50}$ | $IR_{D\geq 75}$ | $IR_{d\leq 10}$ | $IR_{d\leq 5}$ | Score |
| Config 1: η = 0,0025 ; B = 2 | 79,34±0,94 | 89,99±1,49 | 79,20±1,62 | 83,13±1,53 | 64,50±1,32 | 79,27 |
| Config 2: η = 0,0025 ; B = 1 | 74,90±6,55 | 84,48±8,32 | 73,34±7,86 | 77,84±7,79 | 60,09±6,07 | 74,42 |
| Config 3: η = 0,0050 ; B = 2 | 79,85±1,14 | 90,38±1,75 | 79,75±1,89 | 83,45±1,45 | 64,87±1,64 | 79,73 |
| Config 4: η = 0,0075 ; B = 2 | 79,94±0,78 | 90,28±1,19 | 80,28±1,44 | 83,17±1,16 | 65,13±1,60 | 79,83 |
| Config 5 : η = 0,01 ; B = 2 | 78,82±1,75 | 89,49±1,860 | 78,10±3,61 | 81,97±2,46 | 62,67±2,47 | 78,44 |

## ANNEXE II

## ANALYSES STATISTIQUES DES DIFFÉRENCES SIGNIFICATIVES ENTRE LES CNNS POUR LES CINQ MÉTRIQUES

Tableau-A II 1

| $IR_{D\geq 50}$ | | |
|---|---|---|
| **Test de Kruskal-Wallis** | *Rang moyen* | Taille de l'échantillon : 116 |
| *DetectoRS* | 272,15 | |
| *RetinaMask* | 241,72 | |
| *MS R-CNN* | 259,82 | |
| *YOLACT* | 156,32 | Probabilité = 0 |

Tableau-A II 2

| $IR_{D\geq 50}$ | | | |
|---|---|---|---|
| **Intervalles de Bonferroni à 95%** | *Différence* | *Différence significative ?* | |
| *DetectoRS – RetinaMask* | 30,43 | NON | |
| *DetectoRS – MS R-CNN* | 12,33 | NON | |
| *DetectoRS – YOLACT* | 115,83 | OUI | Limites inférieure et supérieure = 46,45 |
| *RetinaMask – MS R-CNN* | -18,10 | NON | |
| *RetinaMask – YOLACT* | 85,41 | OUI | |
| *MS R-CNN – YOLACT* | 103,50 | OUI | |

Tableau-A II 3

| $IR_{D\geq 75}$ | | |
|---|---|---|
| **Test de Kruskal-Wallis** | *Rang moyen* | Taille de l'échantillon : 116 |
| *DetectoRS* | 287,92 | |
| *RetinaMask* | 256,38 | |
| *MS R-CNN* | 251,36 | |
| *YOLACT* | 134,34 | Probabilité = 0 |

Tableau-A II 4

| $IR_{D\geq 75}$ | | | |
|---|---|---|---|
| **Intervalles de Bonferroni à 95%** | *Différence* | *Différence significative ?* | |
| *DetectoRS – RetinaMask* | 31,53 | NON | |
| *DetectoRS – MS R-CNN* | 36,56 | NON | |
| *DetectoRS – YOLACT* | 153,58 | OUI | Limites inférieure et supérieure = 46,45 |
| *RetinaMask – MS R-CNN* | 5,03 | NON | |
| *RetinaMask – YOLACT* | 122,04 | OUI | |
| *MS R-CNN – YOLACT* | 117,02 | OUI | |

Tableau-A II 5

| $IR_{d\leq10}$ | | |
|---|---|---|
| **Test de Kruskal-Wallis** | *Rang moyen* | Taille de l'échantillon : 116 |
| *DetectoRS* | 288,24 | |
| *RetinaMask* | 257,18 | |
| *MS R-CNN* | 263,09 | |
| *YOLACT* | 121,49 | Probabilité = 0 |

Tableau-A II 6

| $IR_{d\leq10}$ | | | |
|---|---|---|---|
| **Intervalles de Bonferroni à 95%** | *Différence* | *Différence significative ?* | |
| *DetectoRS – RetinaMask* | 31,06 | NON | Limites inférieure et supérieure = 46,45 |
| *DetectoRS – MS R-CNN* | 25,16 | NON | |
| *DetectoRS – YOLACT* | 166,75 | OUI | |
| *RetinaMask – MS R-CNN* | -5,91 | NON | |
| *RetinaMask – YOLACT* | 135,69 | OUI | |
| *MS R-CNN – YOLACT* | 141,60 | OUI | |

Tableau-A II 7

| $IR_{d\leq5}$ | | |
|---|---|---|
| **Test de Kruskal-Wallis** | *Rang moyen* | Taille de l'échantillon : 116 |
| *DetectoRS* | 314,37 | |
| *RetinaMask* | 263,44 | |
| *MS R-CNN* | 242,82 | |
| *YOLACT* | 109,37 | Probabilité = 0 |

Tableau-A II 8

| $IR_{d\leq5}$ | | | |
|---|---|---|---|
| **Intervalles de Bonferroni à 95%** | *Différence* | *Différence significative ?* | |
| *DetectoRS – RetinaMask* | 50,93 | OUI | Limites inférieure et supérieure = 46,45 |
| *DetectoRS – MS R-CNN* | 71,55 | OUI | |
| *DetectoRS – YOLACT* | 205,00 | OUI | |
| *RetinaMask – MS R-CNN* | 20,62 | NON | |
| *RetinaMask – YOLACT* | 154,07 | OUI | |
| *MS R-CNN – YOLACT* | 133,46 | OUI | |

Tableau-A II 9

| **Dice** | | |
|---|---|---|
| **Test de Kruskal-Wallis** | *Rang moyen* | Taille de l'échantillon : 116 |
| *DetectoRS* | 288,14 | |
| *RetinaMask* | 254,60 | |
| *MS R-CNN* | 243,14 | |
| *YOLACT* | 144,13 | Probabilité = 0 |

Tableau-A II 10

| **Dice** | | | |
|---|---|---|---|
| **Intervalles de Bonferroni à 95%** | *Différence* | *Différence significative ?* | |
| *DetectoRS – RetinaMask* | 33,54 | NON | Limites inférieure et supérieure = 46,45 |
| *DetectoRS – MS R-CNN* | 45,0 | NON | |
| *DetectoRS – YOLACT* | 144,01 | OUI | |
| *RetinaMask – MS R-CNN* | 11,46 | NON | |
| *RetinaMask – YOLACT* | 110,47 | OUI | |
| *MS R-CNN – YOLACT* | 99,01 | OUI | |

## ANNEXE III

## EXEMPLES D'ANOMALIES CORRIGÉES SUR DES IMAGES PAR POST-TRAITEMENT ANATOMIQUEMENT CONTRAIGNANT

Sur les deux cas, l'image à gauche présente le résultat avec le post-traitement naïf et l'image à droite est le résultat corrigé par le post-traitement anatomiquement contraignant. Dans le premier cas, T1 et T2 sont superposées, T4 est manquée. Dans le second cas, L1 et L2 sont superposées, L5 est manquée.

Figure-A III-1

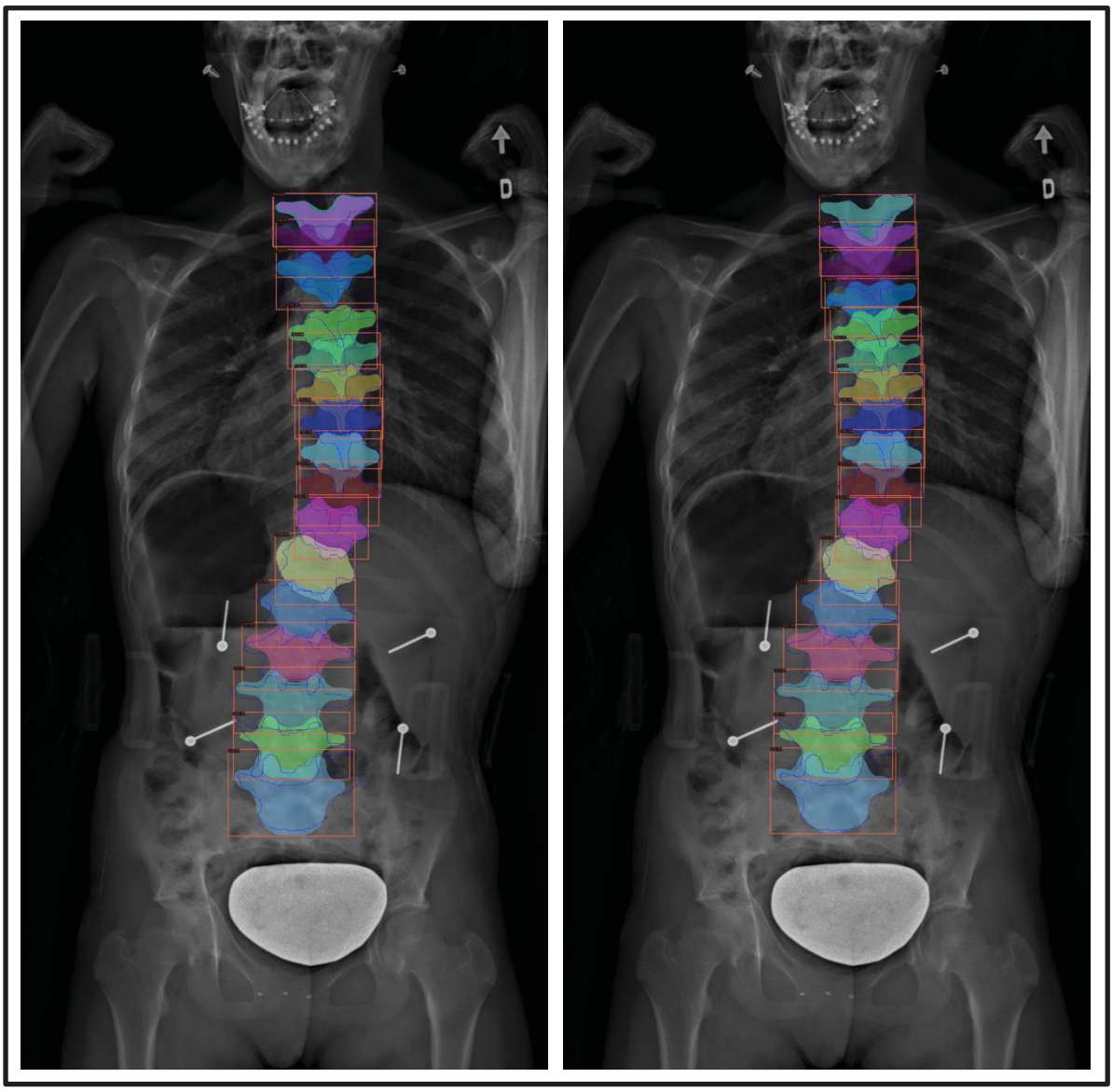
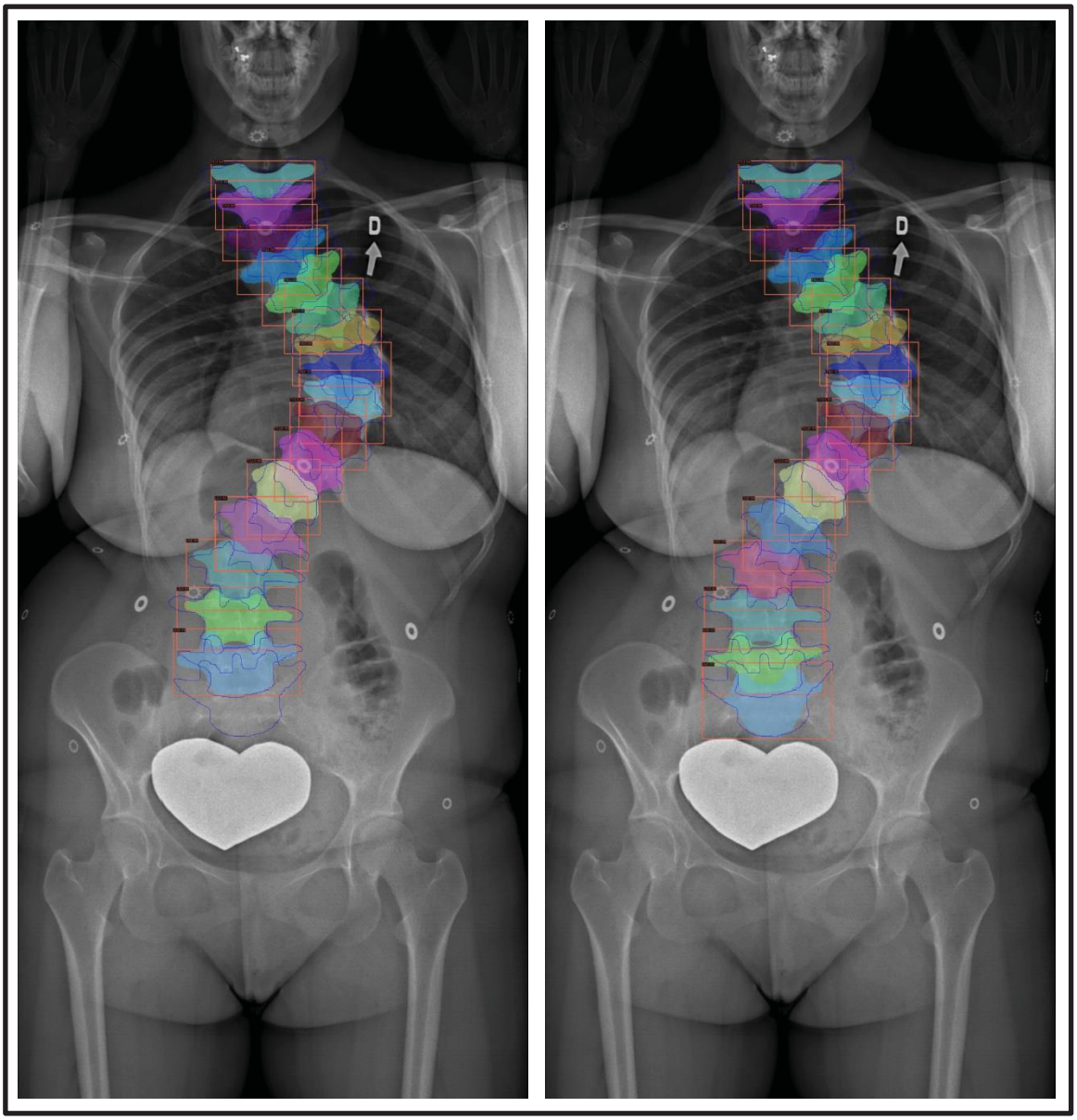


Figure-A III-2



## LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES


Abdulla, W. (2018). Splash of Color: Instance Segmentation with Mask R-CNN and TensorFlow. Repéré à https://engineering.matterport.com/splash-of-color-instance-segmentation-with-mask-r-cnn-and-tensorflow-7c761e238b46

Amidi, A. & Amidi, S. (2021). Convolutional Neural Networks cheatsheet [Notes de cours Repéré à https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks#layer

Asgari Taghnaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J. & Hamarneh, G. (2021). Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, *54*(1), 137-178. doi: 10.1007/s10462-020-09854-1

Aubert, B. (2020). *Reconstruction 3D automatique de la colonne vertébrale à partir de radiographies biplanaires EOS*. (Thèse de doctorat, École de technologie supérieure, Montréal QC). Repéré à https://espace.etsmtl.ca/id/eprint/2716/

Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures ». Dans Montavon, G., Orr, G. B. & Müller, KR (Éds), *Neural Networks: Tricks of the Trade. Lecture Notes in Computer* (Vol. 7700, pp. 437-478). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-35289-8_26

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). YOLACT: Real-Time Instance Segmentation ». Dans *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9156-9165). doi: 10.1109/ICCV.2019.00925

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8,*3-62

Brownlee, J. (2017). What is the Difference Between Test and Validation Datasets?. Repéré à https://machinelearningmastery.com/difference-test-validation-datasets

Brownlee, J. (2019)a. What is the Difference Between a Parameter and a Hyperparameter?. Repéré à https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/

Brownlee, J. (2019)b. How to use Learning Curves to Diagnose Machine Learning Model Performance. Repéré à https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Brownlee, J. (2019)c. How to Control the Stability of Training Neural Networks With the Batch Size. Repéré à https://machinelearningmastery.com/how-to-control-the-speed-and-stability-of-training-neural-networks-with-gradient-descent-batch-size/

Brownlee, J. (2020). How to Choose Loss Functions When Training Deep Learning Neural Networks. Repéré à https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/

Cai, Z. & Vasconcelos, N. (2018). Cascade R-CNN: Delving into High Quality Object Detection. *2018 IEEE/CVF International Conference on Computer Vision and Pattern Recognition* (pp. 6154-6162). doi: 10.1109/CVPR.2018.00644

Chan, L., Hosseini, M.S. & Plataniotis, K.N. (2021). A Comprehensive Analysis of Weakly-Supervised Semantic Segmentation in Different Image Domains. *International Journal of Computer Vision*, *129*(2), 361-384. doi: 10.1007/s11263-020-01373-4

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., … Lin, D. (2019). Hybrid Task Cascade for Instance Segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4969-4978). doi: 10.1109/CVPR.2019.00511

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., … Lin, D. (2019). MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155.* Repéré à http://arxiv.org/abs/1906.07155

Chen, LC., Hermans, A., Papandreou, G., Schroff, F., Wang, P. & Adam, H. (2017). MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. *2018 IEEE/CVF International Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR.2018.00422

Chen, LC., Hermans, A., Papandreou, G., Schroff, F., Wang, P. & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint arXiv:1706.05587*. Repéré à http://arxiv.org/abs/1706.05587

Chen, Y., Gao, Y., Li, K., Zhao, L. & Zhao, J. (2019). Vertebrae Identification and Localization Utilizing Fully Convolutional Networks and a Hidden Markov Model. *IEEE transactions on medical imaging*, *39*(2), 387-399. doi: 10.1109/TMI.2019.2927289

Choudhry, M. N., Ahmad, Z. & Verma, R. (2016). Adolescent Idiopathic Scoliosis. *The open orthopaedics journal* (Vol. 10, pp. 143-154). doi: 10.2174/1874325001610010143

Dhillon, A. & Verma, G. (2019). Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence* (Vol. 9). doi: 10.1007/s13748-019-00203-0

Ekkis, A. (2018). What is the difference between transfer learning and fine tuning?. Repéré à https://www.quora.com/What-is-the-difference-between-transfer-learning-and-fine-tuning

EOS imaging. (2021). *Système EOS*. Repéré à https://www.eos-imaging.com/fr/notre-expertise/solutions-imagerie/systeme-eos

Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T.L. (2017). Machine Learning for Medical Imaging. *RadioGraphics*, *37*(2), 505-515. doi: 10.1148/rg.2017160130

Fu, CY., Shvets, M. & Berg, A C. (2019). RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free. *arXiv preprint arXiv:1901.03353*. Repéré à http://arxiv.org/abs/1901.03353

Fu, CY., Shvets, M. (2019). retinamask. Repéré à https://github.com/chengyangfu/retinamask

Fu, H., Cheng, J., Xu, Y., Wong, D. W. K., Liu, J. & Cao, X. (2018). Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation. *IEEE Transactions on Medical Imaging*, *37*(7), 1597-1605. doi: 10.1109/TMI.2018.2791488

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V. & Garcia-Rodriguez, J. (2017). A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv preprint arXiv:1704.06857*. Repéré à http://arxiv.org/abs/1704.06857

Garyfallos, S., Biseda, B. & Khan, M. (2019). NIH-Chest-X-rays-Classification. Repéré à https://github.com/paloukari/NIH-Chest-X-rays-Classification

Girshick, R. (2015). Fast R-CNN. Dans *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440-1448. doi: 10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2016). Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(1), 142-158. doi: 10.1109/TPAMI.2015.2437384

Glocker, Ben, et al. Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-33454-2_73

Hafiz, A. M. et Bhat, G. M. (2020). A Survey on Instance Segmentation: State of the art. *International Journal of Multimedia Information Retrieval*, *9*(3), 171-189. doi: 10.1007/s13735-020-00195-x

Haque, I. R. I. & Neubert, J. (2020). Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, *18*, 100297. doi: 10.1016/j.imu.2020.100297

He, K., Girshick, R. & Dollar, P. (2019). Rethinking ImageNet Pre-Training. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4917-4926. doi: 10.1109/ICCV.2019.00502

He, K., Gkioxari, G., Dollár, P. & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980-2988. doi: 10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. doi: 10.1109/CVPR.2016.90

Hosch, W. L. (2021). machine learning. Repéré à https://www.britannica.com/technology/machine-learning

Huang, Z., Huang, L., Gong, Y., Huang, C., & Wang, X. (2019). Mask Scoring R-CNN. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6402-6411. doi: 10.1109/CVPR.2019.00657

Hui, J. (2018). mAP (mean Average Precision) for Object Detection. Repéré à https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173

Humbert, L., De Guise, J.A., Aubert, B., Godbout, B. & Skalli, W. (2009). 3D reconstruction of the spine from biplanar X-rays using parametric models based on transversal and longitudinal inferences. *Medical Engineering & Physics*, *31*(6), 681-687. doi: 10.1016/j.medengphy.2009.01.003

Illés, T.S., Lavaste, F. & Dubousset, J. F. (2019). The third dimension of scoliosis: The forgotten axial plane. *Orthopaedics & Traumatology: Surgery & Research*, *105*(2), 351-359.doi: 10.1016/j.otsr.2018.10.021

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., … Ng, A. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 590-597. doi: 10.1609/aaai.v33i01.3301590

Jamesmf. (2017). What do the fully connected layers do in CNNs?. Repéré à https://stats.stackexchange.com/q/182122.

Jia, Z., Huang, X., Chang, E. IC. & Xu, Y. (2017). Constrained Deep Weak Supervision for Histopathology Image Segmentation. *IEEE Transactions on Medical Imaging*, *36*(11), 2376-2388. doi: 10.1109/TMI.2017.2724070

Karpathy, A. & Grebu, T. (2021). Convolutional Neural Networks [Notes de cours]. Repéré à https://cs231n.github.io/convolutional-networks

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ben Ayed, I. (2019). Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis*, *54*, 88-99. doi: 10.1016/j.media.2019.02.009

Kim, K. C., Cho, H. C., Jang, T. J., Choi, J. M. & Seo, J. K. (2021). Automatic detection and segmentation of lumbar vertebrae from X-ray images for compression fracture evaluation. *Computer Methods and Programs in Biomedicine*, *200*, 105833. doi: 10.1016/j.cmpb.2020.105833

Kim, K. C., Cho, H. C., Jang, T. J., Choi, J. M. & Seo, J. K. (2020). Automation of Spine Curve Assessment in Frontal Radiographs Using Deep Learning of Vertebral-Tilt Vector. *IEEE Access*, *8*, 84618-84630. doi: 10.1109/ACCESS.2020.2992081

Konieczny, M. R., Senyurt, H. & Krauspe, R. (2013). Epidemiology of adolescent idiopathic scoliosis. *Journal of Children's Orthopaedics*, *7*(1), 3-9. doi: 10.1007/s11832-012-0457-4

Kónya, S., Natarajan, T. R. S., Allouch, H., Nahleh, K. A., Dogheim, O. Y. & Boehm, H. (2021). Convolutional neural network-based automated segmentation and labeling of the lumbar spine X-ray. *Journal of Craniovertebral Junction & Spine*, *12*(2), 136-143. doi: 10.4103/jcvjs.jcvjs_186_20

Kornblith, S., Shlens, J. & Le, Q. C. (2019). Do Better ImageNet Models Transfer Better?. *arXiv preprint arXiv:1805.*08974. Repéré à http://arxiv.org/abs/1805.08974

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583-621. doi: 10.2307/2280779

Labelle, H., Aubin, CE., Jackson, R., Lenke, L., Newton, P. & Parent, S. (2011). Seeing the Spine in 3D: How Will It Change What We Do?. *Journal of Pediatric Orthopaedics*, *31*, S37-S45. doi: 10.1097/BPO.0b013e3181fd8801

Lamarre, ME. (2008). *Conception d'une méthode standardisée pour évaluer la flexibilité de la colonne vertébrale*. (Mémoire à la maitrise, École de technologie supérieure, Montréal, QC). Repéré à http://espace.etsmtl.ca/id/eprint/142

Lambert, F. M., Malinvaud, D., Glaunès, J., Bergot, C., Straka, H. & Vidal, PP. (2009). Vestibular asymmetry as the cause of idiopathic scoliosis: a possible answer from Xenopus. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(40), 12477-12483. doi: 10.1523/JNEUROSCI.2583-09.2009

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. doi: 10.1109/5.726791

Lessmann, N., van Ginneken, B., de Jong, P. A. & Išgum, I.. (2019). Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical Image Analysis*, *53*, 142-155. doi: 10.1016/j.media.2019.02.005

Li, FF., Johnson, J. & Yeung, S. [Stanford University School of Engineering]. (2017). *Lecture 11: Detection and Segmentation* [Vidéo en ligne]. Repéré à https://www.youtube.com/watch?v=nDPWywWRIRo

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936-944. doi: 10.1109/CVPR.2017.106

Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999-3007. doi: 10.1109/ICCV.2017.324

Lin, T. Y., Maire, M., Belongie,S., Hays, J., Perona, P., Ramanan, D., … Zitnick, L. C. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*, 740-755. Cham : Springer International Publishing.

Long, J., Shelhamer, E. & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *arXiv preprint arXiv:1411.4038*. Repéré à http://arxiv.org/abs/1411.4038

Machida, M. (1999). Cause of Idiopathic Scoliosis. *Spine*, *24*(24), 2576.

Marius, P., Balas, V., Perescu-Popescu, L. & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, *8*.

Márquez-Neila, P., Salzmann, M. & Fua, P. (2017). Imposing Hard Constraints on Deep Networks: Promises and Limitations. *arXiv preprint arXiv:1706.02025.* Repéré à http://arxiv.org/abs/1706.02025

Massa, F. & Girshick, R. (2018). *maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch*. Repéré à https://github.com/facebookresearch/maskrcnn-benchmark

McDonald, J. H. (2015). Multiple comparisons. Repéré à http://www.biostathandbook.com/multiplecomparisons.html

Moulin, D. (2014). *Développement et évaluation d'une méthode de mesure de la flexibilité du rachis scoliotique*. (Mémoire à la maitrise, École de technologie supérieure, Montréal QC). Repéré à http://espace.etsmtl.ca/id/eprint/1380

Murphy, K. P. (2012). Probability. *Machine learning: a probabilistic perspective*. Cambridge, MA : MIT Press.

Negrini, S., Aulisa, A. G., Aulisa, L., Circo, A. B., de Mauroy, J. C., Durmala, J., … Zaina, F. (2012). 2011 SOSORT guidelines: Orthopaedic and Rehabilitation treatment of idiopathic scoliosis during growth. *Scoliosis*, *7*(1), 3. doi :10.1186/1748-7161-7-3

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. *2017 15th International Conference on ICT and Knowledge Engineering (ICT KE)*,1-6. doi: 10.1109/ICTKE.2017.8259629

Pathak, D., Krähenbühl, P. & Darrell, T. (2015). Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*,1796-1804. doi: 10.1109/ICCV.2015.209

Payer, C., Štern, D., Bischof, H. & Urschler, M. (2019). Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical Image Analysis*, *54*, 207-219. doi: 10.1016/j.media.2019.03.007

Payer, C., Štern, D., Bischof, H. & Urschler, M. (2021). Coarse to Fine Vertebrae Localization and Segmentation with SpatialConfiguration-Net and U-Net. *15th International Conference on Computer Vision Theory and Applications*, 5, 124-133. doi: 10.5220/0008975201240133

Pomero, V., Mitton, D., Laporte, S., de Guise, J. A. & Skalli, W. (2004). Fast accurate stereoradiographic 3D-reconstruction of the spine using a combined geometric and statistic model. *Clinical Biomechanics*, *19*(3), 240-247. doi: 10.1016/j.clinbiomech.2003.11.014

Pramoditha, R. (2020). *k-fold cross-validation explained in plain English*. Repéré à https://towardsdatascience.com/k-fold-cross-validation-explained-in-plain-english-659e33c0bc0

Pröve, P. L. (2017). An Introduction to different Types of Convolutions in Deep Learning. Repéré à https://towardsdatascience.com/types-of-convolutions-in-deep-learning-717013397f4d

Qiao, S., Chen, L. C. & Yuille, A. (2021). DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,10208-10219. doi: 10.1109/CVPR46437.2021.01008

Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. (2019). Transfusion: Understanding Transfer Learning with Applications to Medical Imaging. *arXiv preprint arXiv:1902.07208.* Repéré à https://arxiv.org/abs/1902.07208

Rasoulian, A., Rohling, R., & Abolmaesumi, P. (2013). Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+ pose model. *IEEE transactions on medical imaging*, 32(10), 1890-1900. doi: 10.1109/TMI.2013.2268424

Redmon, J., Divvala, S., Girshick, R & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788. doi: 10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R. & Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137-1149. doi: 10.1109/TPAMI.2016.2577031

Romero, M., Interian, Y., Solberg, T. & Valdes, G. (2020). Targeted transfer learning to improve performance in small medical physics datasets. *Medical Physics*, *47*(12), 6246-6256. doi: 10.1002/mp.14507

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*., 234-241. Cham: Springer International Publishing.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. doi: 10.1037/h0042519

Shapiro, S. S. & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591-611. doi: 10.2307/2333709

scikit-learn developers. (2021). Cross-validation: evaluating estimator performance. In *scikit-learn*. Repéré à https://scikit-learn.org/stable/modules/cross_validation.html

Sekuboyina, A., Bayat, A., Husseini, M., Löffler, M., Rempfler, M., Kukačka, J., …, Kirschke, J. (2021). VerSe: A Vertebrae Labelling and Segmentation Benchmark. *Medical Image Analysis, 73,* 102166. doi: 10.1016/j.media.2021.102166

Siddique, N., Paheding, S., Elkin, C. P. & Devabhaktuni, V. (2021). U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, *9*, 82031-82057. doi: 10.1109/ACCESS.2021.3086020

Simonyan, K. & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Dans Bengio, Y. & LeCun, Y. (Éds) *3rd International Conference on Learning Representations*. Repéré à http://arxiv.org/abs/1409.1556

Soviany, P. & Ionescu, R. D. (2018). Optimizing the Trade-Off between Single-Stage and Two-Stage Deep Object Detectors using Image Difficulty Prediction. *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 209-214. doi: 10.1109/SYNASC.2018.00041

Taud, H. et J.F. Mas. (2018). Multilayer Perceptron (MLP). Dans Olmedo, C., Teresa, M., Paegelow, M., Mas, J. F. & Escobar, F. *Geomatic Approaches for Modeling Land Change Scenarios*, 451-455. Cham: Springer International Publishing. doi: 10.1007/978-3-319-60801-3_27

Tsai, Y. C., Chen, J. H. & Wang, J. J. (2018). Predict Forex Trend via Convolutional Neural Networks. *Journal of Intelligent Systems*, *29*(1), 941-958. doi: 10.1515/jisys-2018-0074

Walsh, J., O' Mahony, N., Campbell, S., Carvalho, A., Krpalkova, l., Velasco-Hernandez, G., Harapanahalli, S. & Riordan, D. (2019). *Deep Learning vs. Traditional Computer Vision*. doi :10.1007/978-3-030-17795-9_10

Wang, R., Yi Voon, J. H., Ma, D., Dabiri, S., Popuri, K. & Beg, M. F. (2021). Vertebra Segmentation for Clinical CT Images Using Mask R-CNN. *8th European Medical and Biological Engineering Conference*, 1156-1165. Cham: Springer International Publishing.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 3462-3471. doi: 10.1109/CVPR.2017.369

Weinstein, S. L., Dolan, L. A., Cheng, J. CY., Danielsson, A. & Morcuende, J. A. (2008). Adolescent idiopathic scoliosis. *The Lancet*, *371*(9623) 1527-1537. doi: 10.1016/S0140-6736(08)60658-3

Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987-5995. doi: 10.1109/CVPR.2017.634

Yan, Z., Liang, J., Pan, W., Li, J. & Zhang, C. (2017). Weakly- and Semi-Supervised Object Detection with Expectation-Maximization Algorithm. *arXiv preprint arXiv:1702.8740.* Repéré à https://arxiv.org/abs/1702.08740

Yang, Z., Skalli, W., Vergari, C., Angelini, E. D. & Gajny, L. (2019). Automated Spinal Midline Delineation on Biplanar X-Rays Using Mask R-CNN. Tavares, J. M. R. S. & Natal Jorge, R. M. (Éds) *VipIMAGE* 2019 (Vol. 34, pp. 307-316). Cham : Springer International Publishing. doi: 10.1007/978-3-030-32040-9_32

Yohanandan, S. (2020). mAP (mean Average Precision) might confuse you!. Repéré à https://towardsdatascience.com/map-mean-average-precision-might-confuse-you-5956f1bfa9e2

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014). How transferable are features in deep neural networks?. *arXiv preprint arXiv:1411.1792.* Repéré à http://arxiv.org/abs/1411.1792

Zavalina, M., Jain, P., Pearl, A. & Kollmar, D. (2020). *Evolution and Uses of CNNs and Why Deep Learning?*. Repéré à https://atcold.github.io/pytorch-Deep-Learning/en/week01/01-2

Zhang, H. & Hong, X. (2019). Recent progresses on object detection: a brief review. *Multimedia Tools and Applications*, *78*(19), 27809-27847. doi: 10.1007/s11042-019-07898-2

Zhang, R., Xiao, X., Liu, Z., Li, Y. & Li, S. (2020). MRLN: Multi-Task Relational Learning Network for MRI Vertebral Localization, Identification, and Segmentation. *IEEE Journal of Biomedical and Health Informatics*, vol. *24*(10), 2902-2911. doi: 10.1109/JBHI.2020.2969084

Zhao, H., Shi, J., Qi,X., Wang, X. & Jia, J. (2017). Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230-6239. doi: 10.1109/CVPR.2017.660