Apprentissage profond pour l'identification automatique de l'indice de Risser

par

Houda KADDIOUI

MÉMOIRE PAR ARTICLE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE AVEC MÉMOIRE EN TECHNOLOGIE DE LA SANTÉ
M.Sc.A.

MONTRÉAL, LE 4 JANVIER 2019

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Luc Duong, PhD, directeur de recherche
Département de génie logiciel et des technologies de l'information à l'École de Technologie Supérieure

Mme. Sylvie Raté, PhD, Présidente du Jury
Département de génie logiciel et des technologies de l'information à l'École de Technologie Supérieure

M. Éric Wagnac, PhD, membre du jury
Département de génie mécanique à l'École de Technologie Supérieure

Dr. Guy Grimard, MD, Examinateur Externe
Département de chirurgie, CHU Sainte-Justine

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

LE 12 DÉCEMBRE 2018

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

À mon directeur de recherche, Prof. Luc Duong : mes sincères remerciements pour la confiance que vous m'avez témoignée en me confiant ce travail. Je vous remercie pour votre bienveillance, pour tout le temps que vous m'avez consacré. Vous m'avez poussé à dépasser mes limites et avez cru en moi. Merci pour vos conseils judicieux et vos remarques hors pairs. Je vous remercie pour votre engagement et votre dévouement pour vos étudiants et pour faire avancer la recherche au bénéfice des patients.

À Prof. Sylvie Raté : merci de nous faire l'honneur de présider ce jury. Je vous remercie également de continuer à lutter pour le bien de vos étudiants. Merci également pour vos commentaires judicieux lors des diverses présentations de laboratoire. Vous êtes une inspiration pour les futures femmes en génie. Je vous exprime ma gratitude.

Je souhaite remercier sincèrement le Professeur Eric Wagnac d'accepter de faire parti de ce jury et d'évaluer ce travail.

Mes remerciements au Dr. Guy Grimard d'avoir accepté de participer à ce jury. Ce projet n'aurait pas abouti sans votre bienveillance et conseils. Merci de m'avoir guidée aux stages préliminaires de ce projet et d'avoir pris le temps de répondre à mes questions, malgré vos occupations et votre temps précieux.

Je tiens à remercier spécialement Julie Joncas, qui a proposé la problématique de recherche. Son aide précieuse à grandement contribué à l'aboutissement de ce projet. Une reconnaissance particulière est due aux Drs Labelle, Parent, Nahle, Nault et Chemaly, et enfin à Christian Bellefleur pour leur implication dans les différentes étapes de la rédaction article.

Pour finir, je souhaiterai rendre hommage aux collègues des laboratoires LIVE et LINCS (A3440). À ceux qui sont passés très vite mais nous ont un peu changés, à ceux qui sont là tous les jours, à Rémi, Fariba, Gerardo, Edgar et Atefeh pour toujours prendre le temps de discuter de nos projets et émettre des suggestions. Aux personnes qui comptent le plus pour moi. Elles se reconnaîtront.

Apprentissage profond pour l'identification automatique de l'indice de Risser

Houda KADDIOUI

RÉSUMÉ

La scoliose idiopathique de l'adolescent (SIA) est une déformation de la colonne vertébrale d'origine inconnue. C'est une pathologie fréquente qui touche 1 à 3% des adolescents, avec une prédominance féminine. Le traitement en SIA dépends essentiellement du type de courbure que le patient présente. Le type de déformation influence la stratégie de prise en charge qui va de la simple observation aux chirurgies réparatrices très invasives dans les cas de déformations plus sévères. Cependant, pour les déformations peu sévères au moment de la consultation, la décision d'opérer dépendra du potentiel de progression de la courbure prévue par le chirurgien. Les meilleurs indicateurs de la progression sont le potentiel de croissance et la vitesse de croissance. Ces derniers dépendent de la maturité osseuse au niveau du bassin, et plus précisément de la croissance de l'apophyse iliaque. La maturité osseuse est déterminée chez les patients grâce à l'indice de Risser. Cette évaluation peut cependant s'avérer difficile et est sujette à une variabilité intra et inter-évaluateur. Cette étude vise à développer une méthode automatique, fiable et reproductible pour l'évaluation de l'indice de Risser grâce à des méthodes d'apprentissage profond.

Un réseau de neurones convolutif a été développé pour automatiser la lecture de l'indice de Risser sur des radiographies conventionnelles. Le réseau a ensuite été validé en comparant sa précision à la variabilité inter et intra-observateur de six experts.

L'accord global entre les observateurs a été jugé modéré, avec un coefficient kappa de 0,60 pour le groupe d'observateurs expérimentés et un accord de 74,50%. La méthode de classification automatique a montré un coefficient kappa de 0,72, ce qui est un accord fort avec la vérité étalon, et une précision globale de 78,00%.

Ceci est la première étude utilisant l'apprentissage profond pour la détermination automatique de l'indice de Risser. Ce travail apporte une nouvelle méthode pour la normalisation de la classification de Risser, et fournit des informations supplémentaires dans l'évaluation de la maturité osseuse à partir de radiographies.

Mots-clés : Apprentissage profond, Scoliose idiopathique de l'adolescent (SIA), rachis, Indice de Risser, réseau de neurones convolutif

Automatic Risser Grade Assessment Using Deep Learning

Houda KADDIOUI

ABSTRACT

The Risser grade is a widely used indicator of bone maturity in the management panel of Adolescent Idiopathic Scoliosis (AIS). The best predictors of curve progression are the growth potential and kinetics, which both depend on the bone's maturity. However, bone maturity assessment from radiographs is challenging and is subject to intra and inter-observer variability. This study aims at developing an automatic, reliable and reproducible method for the assessment of the Risser's grade using deep learning.

A convolutional neural network was trained to automatically grade conventional radiographs according to the Risser method. A total of 1830 posteroanterior radiographs of AIS patients were retrospectively collected and graded using the American Risser's definition. Each radiograph was pre-processed and cropped to include the entire pelvis region. The network was then validated by comparing its accuracy against the inter and intra-observer variability of six trained graders from our institution.

Overall agreement between observers was fair, with a kappa coefficient of 0.60 for the experienced graders and an agreement of 74.50%. The automatic grading method obtained a kappa coefficient of 0.72, which is a substantial agreement with the ground truth, and an overall accuracy of 78.00%.

This is the first study using deep learning for automatic assessment of the Risser grade. This work may provide a new method for standardization of Risser grading, and additional insights in the assessment of bone maturity from radiographs.

Keywords: Deep learning, Adolescent idiopathic scoliosis (AIS), spine, Risser stage, Convolutional neural networks (CNNs).

TABLE DES MATIÈRES

Page

LISTE DES FIGURES

INTRODUCTION

La scoliose idiopathique de l'adolescent (SIA) est une déformation de la colonne vertébrale dont la prévalence est estimée à 1-3% des adolescents, avec une prédominance féminine. L'origine de la SIA est encore inconnue, mais de nombreux auteurs se sont intéressés à son étiologie. En effet, (Herman *et al.*, 1985) ont démontré l'implication du système nerveux central dans les troubles posturaux dont la SIA. L'origine génétique est également supportée notamment par l'étude d'une cohorte de jumeaux danoise montrant une prévalence plus élevée chez les jumeaux monozygotes (Simony *et al.*, 2016), ou encore l'étude effectuée par Génome Québec qui retrouve 34 gènes candidats (Gorman *et al.*, 2012). Il faut noter cependant que la SIA présente une grande hétérogénéité phénotypique et génétique rendant difficile l'identification d'une cause précise. L'état actuel de la science est ainsi en faveur d'une origine multifactorielle.

Le type de déformation influence largement la stratégie de prise en charge. Le traitement de la SIA dépend ainsi essentiellement du type de courbure que le patient présente. Il peut aller de la simple observation à la chirurgie dans les cas de déformations plus sévères. Cependant, pour les déformations peu sévères au moment de la consultation, la décision d'opérer dépendra du potentiel de progression de la courbure prévue par le chirurgien. Au total, parmi tous les patients atteints, 3-9% seront traités parmi lesquels 90% recevront un traitement conservateur par corset et 10% seront opérés (Goldberg *et al.*, 1988). Dans tous les cas, un diagnostic de SIA confirmé peut avoir des impacts majeurs sur la qualité de vie des patients d'un point de vue physique, mais également psychologique et social. Un suivi régulier est donc prescrit, surtout que les patients atteints de SIA sont plus susceptibles de développer des douleurs dorsales, des complications cardiorespiratoires et des troubles psychologiques (Weinstein *et al.*, 2008). L'imagerie par rayons-X est généralement le premier examen radiologique effectué. La qualité des images peut varier selon les institutions, les techniciens ou les patients, mais elle reste l'examen de première intention. Les incidences frontales (postéro antérieure) et latérales doivent inclure le rachis en entier, ainsi que le pelvis et les deux crêtes iliaques (figure 0.1).

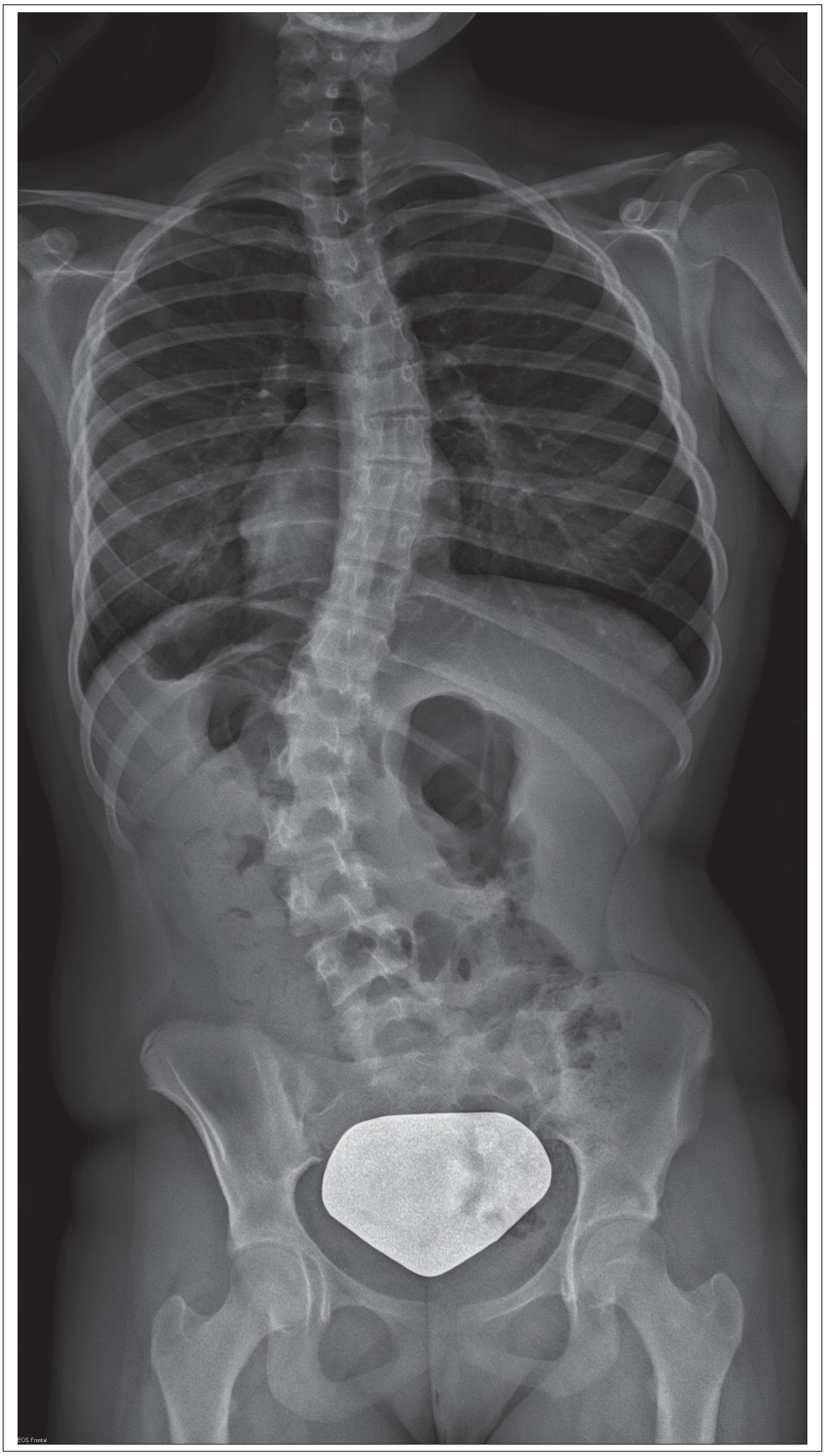

Figure 0.1 Radiographie postéroantérieure montrant une scoliose

L'évaluation et le pronostic de la SIA seront alors déterminés sur les radiographies selon deux critères : la sévérité de la déformation grâce à l'angle de Cobb (angle entre les deux vertèbres limites définissant la courbure du rachis) et la maturité osseuse grâce à l'indice de Risser. La décision d'opérer dépendra du potentiel de progression de la courbure prévue par le chirurgien. Les meilleurs indicateurs de la progression sont le potentiel de croissance et la vitesse de croissance. Ces derniers dépendent de la maturité osseuse mesurée au niveau du bassin, et plus précisément de la croissance de l'apophyse iliaque. La maturité osseuse est déterminée chez les patients scoliotiques grâce à l'indice de Risser. L'indice de Risser est mesuré au niveau de la crête iliaque et reflète la maturité osseuse au niveau de la région du tronc. Il s'agit d'une quantification de l'ossification iliaque en 6 stades ou le stade 0 correspond à l'absence totale d'ossification et le stade 5 à l'ossification complète de l'os. Le potentiel évolutif de la scoliose étant largement dépendant de la maturité osseuse, l'indice de Risser se trouve être une mesure importante dans le cadre de la SIA. En effet, 68% des patients avec un stade 0 ou 1 ont une progression de la courbure au delà de 30 degrés, contre seulement 23% au stade 4. (Horne *et al.*, 2014). L'évaluation de l'indice de Risser peut cependant s'avérer difficile en pratique et est sujette à une variabilité intra et inter-évaluateur.

Comme vu précédemment, la prise en charge de la scoliose se fait en plusieurs étapes. Avec le nombre de patients reçus chaque année, et selon la disponibilités de professionnels dans le domaine de l'orthopédie et de la neurochirurgie pédiatrique, il est possible d'accumuler des délais entre chaque étape. De plus, certaines des composantes de la prise en charge sont sujettes à des variabilité inter-évaluateurs en terme de précision (Ceci sera abordé plus en détail au chapitre suivant). Pour remédier à ces limitations, de nombreux groupes ont tenté d'automatiser le processus de prise en charge en automatisant la segmentation des vertèbres, la mesure de l'angle de Cobb (Sardjono *et al.*, 2013), l'estimation de la sévérité de la courbure (Ramirez *et al.*, 2006) et des algorithmes ont même été développés pour suggérer la stratégie chirurgicale optimale (Phan *et al.*, 2015). Aucune étude à notre connaissance ne s'est intéressée à l'auto-

matisation de la détection de l'indice de Risser, bien que la maturité osseuse soit un élément clé dans la prise en charge de la SIA.

L'objectif de ce mémoire est de développer une méthode automatique, et d'évaluer la précision et la répétabilité pour l'évaluation de l'indice de Risser grâce à des méthodes d'apprentissage profond. Un réseau de neurones convolutif a été développé pour automatiser la lecture de l'indice de Risser sur des radiographies conventionnelles. Le réseau a ensuite été validé en comparant sa précision à la variabilité inter et intra-observateur de six experts.

Le mémoire est organisé comme suit : la revue de littérature introduit le sujet afin d'expliquer les notions de maturité osseuse, d'apprentissage profond ainsi que l'état de l'art actuel en terme d'évaluation de l'âge osseux (Chapitre 1). Un article scientifique a été rédigé et soumis à la revue Radiology : Artificial Intelligence. Cet article, présenté au Chapitre 2, détaille la méthodologie et des résultats obtenus. Ensuite, une discussion revient sur les résultats obtenus, aborde les limites de notre travail ainsi que les perspectives d'avenir et les travaux futurs (Chapitre 3).

CHAPITRE 1

REVUE DE LA LITTÉRATURE

Dans ce chapitre, nous traiterons de l'état de l'art dans la détermination de l'âge osseux dans le contexte de SIA. Dans un premier temps nous exposerons l'anatomie normale de la colonne vertébrale et du bassin (pelvis osseux). Nous expliquerons ensuite les déformations dues à la SIA. Enfin, nous introduirons les méthodes contemporaines d'automatisation de l'âge osseux sur des radiographies et les techniques d'imagerie médicale employées de nos jours.

1.1 Anatomie du rachis et du bassin.

La colonne vertébrale est un élément clé de l'anatomie humaine. Cet enchaînement de vertèbres est agencé pour protéger la moelle épinière, supporter la cage thoracique et les mouvements respiratoires et assurer une stabilité dans les mouvements. La disposition des vertèbres et leurs spécificités mécaniques procurent au rachis des courbures naturelles qui sont la lordose cervicale, cyphose dorsale et lordose lombaire (figure 1.1). Le bassin (figure 1.2) fait suite aux vertèbres lombaires. C'est un massif qui supporte le poids du tronc, s'articule avec les fémurs et protège les structures nobles internes. Le bassin est constitué de trois entités : les deux os coxaux (anciennement appelés os iliaques), le sacrum et le coccyx. Chaque os coxal est lui même composé de trois régions : L'ilium, le pubis et l'ischium. La région d'intérêt pour cette étude est la partie iliaque, et plus précisément la crête iliaque.

1.2 L'âge osseux

Une notion importante à clarifier pour une meilleure compréhension de notre étude est l'âge osseux. En effet, l'âge tel que nous l'utilisons dans la vie courante correspond à l'âge chronologique, c'est-à-dire le nombre d'années écoulées depuis la naissance. Cependant, la définition de la maturité osseuse est basée sur l'âge biologique, dont la référence est la fin de la croissance osseuse. La détermination de l'âge biologique peut se faire cliniquement en observant

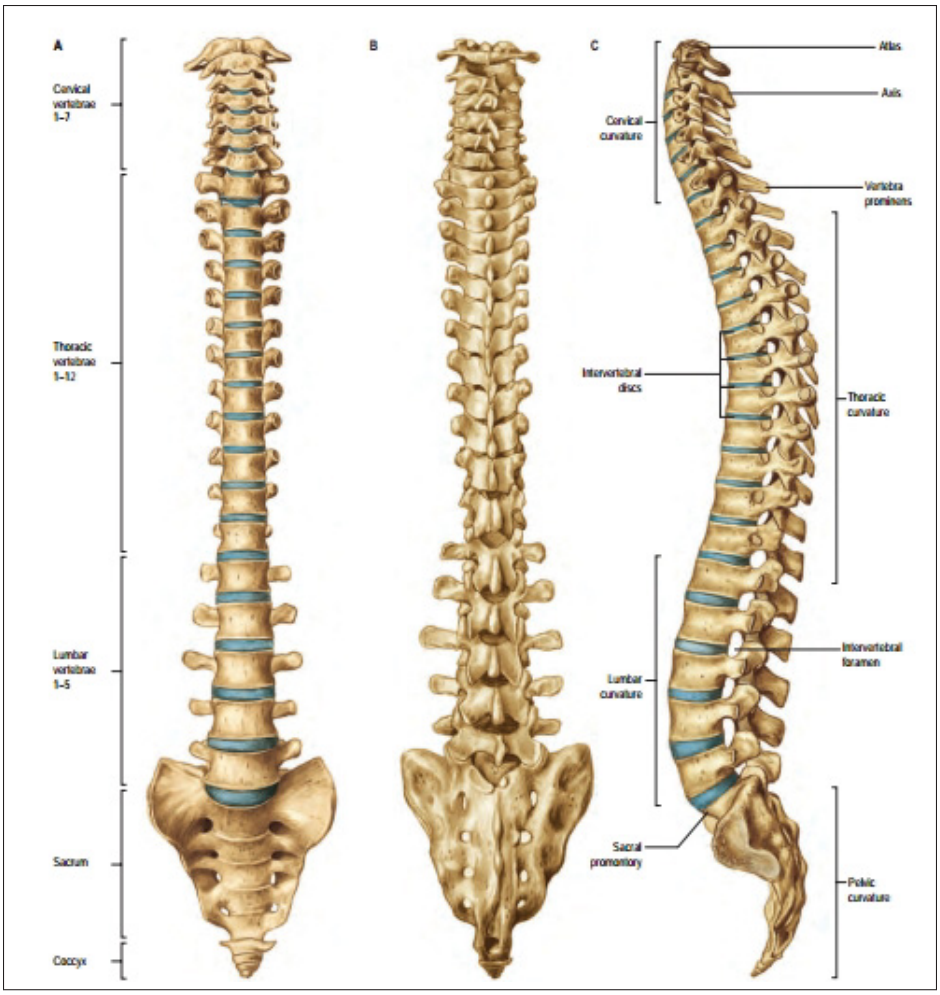

Figure 1.1 Colonne vertébrale, vue antérieure (A) postérieure(B) et latérale(C) Source : Gray's anatomy

les signes physiques de développement pubertaire, pouvant être évalué grâce l'échelle de maturité de Tanner (Marshall & Tanner, 1970)(Marshall & Tanner, 1969). L'imagerie reste tout de même l'outil le plus utilisé pour évaluer l'âge osseux. Parmi les méthodes disponibles, la détermination de l'âge osseux sur les radiographies de poignet est la plus courante dans la pratique médicale. Cependant, c'est une méthode laborieuse, qui demande à comparer 20 régions clés sur les radiologies à un atlas de référence. Dans le cas de la SIA, l'indicateur le plus utilisé est l'indice de Risser.

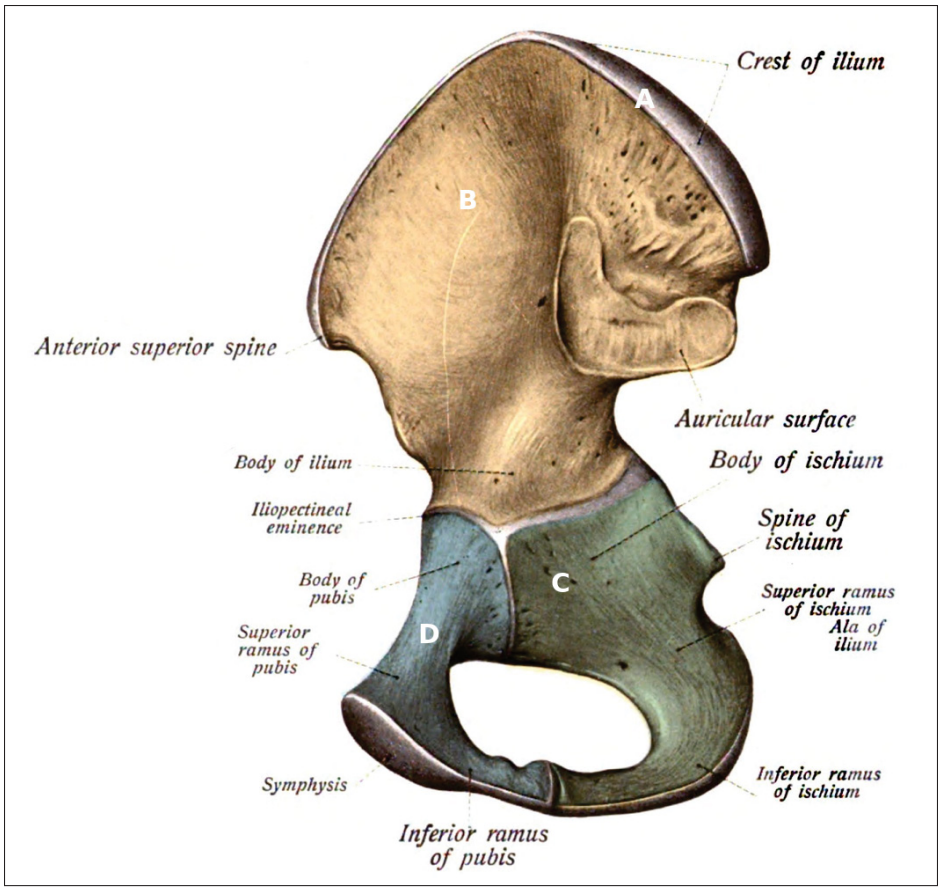

Figure 1.2 Bassin. A : Crête iliaque ; B : Ilium ; C : Ischium ; D : Pubis ; Tirée et modifiée de : Gray's anatomy

En 1958, Risser a introduit une méthode simple pour estimer la maturité osseuse au niveau du bassin. Cette méthode se base sur l'observation de la croissance de la crête iliaque -suivant habituellement un trajet antéro-postérieur- sur des radiographies frontales de bassin, et divise cette progression en différents stades. Il existe actuellement une classification de Risser française et une classification américaine (figure 1.3) (Hacquebord & Leopold, 2012).

La classification française divise la progression de l'ossification en 5 étapes :

- Le stade 0 correspond à l'ossification du premier tiers
- Le stade 1 correspond à l'ossification du second tiers
- Le stade 3 correspond à l'ossification complète de la crête

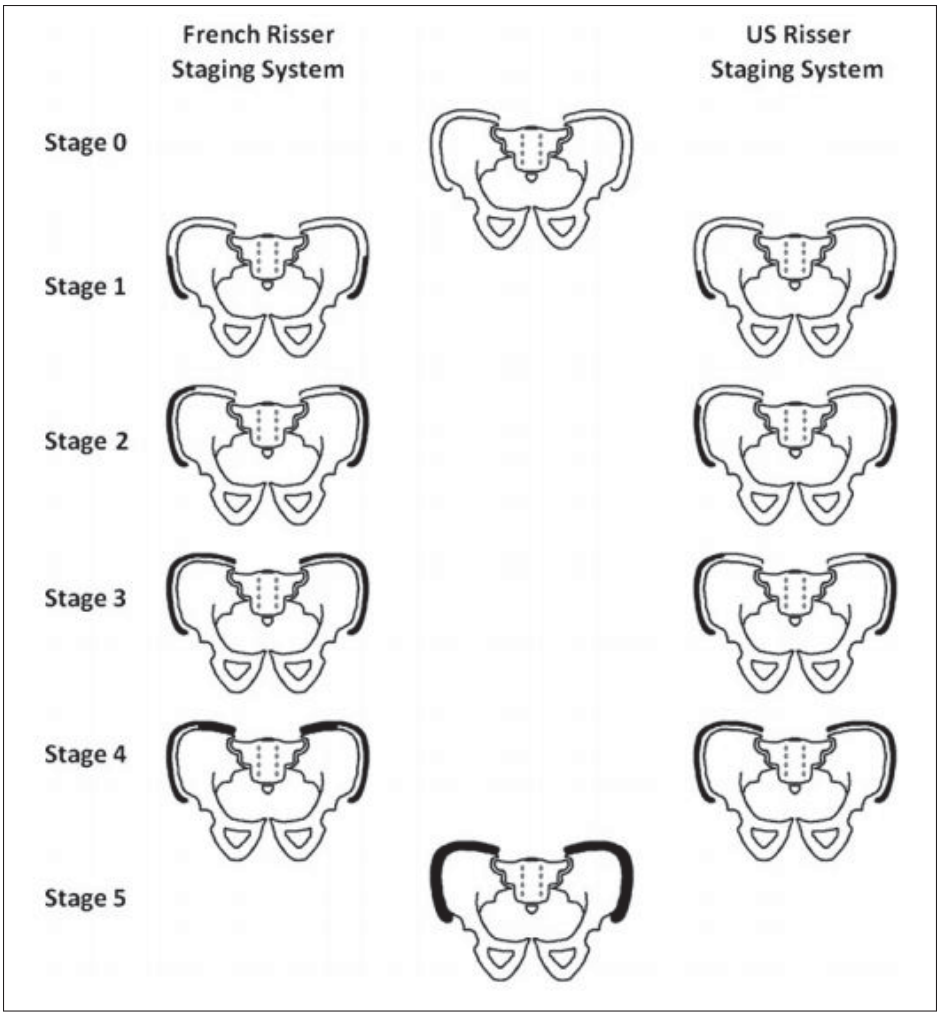

Figure 1.3 Illustration des deux versions de la classification de Risser

- Le stade 4 correspond au début de la fusion de l'os et de la crête iliaque
- Le stade 5 correspond à la fusion totale de la crête sur l'os iliaque.

La classification de Risser américaine divise la progression en 6 étapes :

- Le stade 0 correspond a l'absence d'ossification
- Le stade 1 correspond une ossification à 25%
- Le stade 2 correspond une ossification à 50%
- Le stade 3 correspond une ossification à 75%

- Le stade 4 correspond une ossification à 100%
- Le stade 5 correspond à la fusion de la crête iliaque et de l'os iliaque.

L'avantage de l'indice de Risser est qu'il est déjà disponible sur les radiographies de la colonne qui incluent les fémurs. Les patients n'ont donc pas besoin de subir des irradiations supplémentaires pour évaluer leur maturité osseuse. La prise en charge de la scoliose peut aller de la simple surveillance, aux chirurgies parfois lourdes et invasives. La décision thérapeutique dépend en partie de la maturité osseuse, car le risque évolutif de la scoliose est proportionnel au potentiel de croissance de chaque patient. Cependant, bien que les méthodes de détermination de la maturité osseuse permettent de prendre des décisions thérapeutiques éclairées, elles sont sujettes à des limitations : tout d'abord, il n'existe pas de méthode standard ou de référence pour toutes les applications. De plus, ce sont des méthodes non exactes puisque la décision est basée sur une estimation quantitative visuelle. Ce qui en découle est une variabilité inter-observateur et intra-observateur qui entraîne une disparité dans l'estimation de la maturité osseuse.

1.3 Variabilité inter-observateur et intra-observateur

La limitation majeure en terme de détermination manuelle de l'âge osseux est la variabilité inter-évaluateur et intra-évaluateur qui est démontrée et admise en pratique clinique. Cette variabilité est inhérente à la notion même de croissance osseuse puisqu'il s'agit d'un processus continu, entraînant des changements morphologique que l'humain peut difficilement quantifier. De plus, la rotation du bassin conséquente à la SIA, la variabilité biologique entre les individus et la qualité variable des images à rayon-X sont des facteurs extrinsèques augmentant le risque de variabilité. Cependant, des études ont établi un manque de consensus concernant l'ampleur de cette variabilité dans la détermination du stade de Risser.

En ce qui concerne la variabilité inter-observateurs, Goldberg et al. (Goldberg *et al.*, 1988) ont démontré un accord substantiel avec un coefficient de kappa à 0,80. Dhar et al. (Dhar *et al.*, 1993) ont également démontré un accord de 89,2%.En revanche, des études plus récentes ont montré un accord de 50% tous stades confondus, tandis que Hammond et al, en accord avec

Shuren et al, ont démontré un accord modéré entre chirurgiens orthopédiques et radiologues, avec une différence pouvant aller jusqu'à trois stades entre les évaluateurs (Hammond *et al.*, 2011).

Une autre controverse concerne l'influence de la direction d'acquisition des images radiologiques : (Izumi, 1995) suggèrent que les radiographies antéro-postérieures sont un meilleur reflet de l'ossification iliaque, alors que la majorité des radiologies frontales sont acquises de façon postéro-antérieure pour éviter d'irradier les organes de reproduction. Ils suggèrent que cette différence entre les images entraîne des inexactitudes dans l'évaluation du stade de Risser. Cependant, Reem et al. ont étudié la différence dans la lecture des radiologies antéro-postérieures et postéro-antérieure et démontrent que le stade de Risser se mesure de façon acceptable (Hammond *et al.*, 2011; Reem *et al.*, 2009; Sabour, 2018; Yang *et al.*, 2014).

Cette variabilité évidente peut avoir un impact important lorsque l'on considère les stratégies thérapeutiques et les résultats. De plus, les images d'un même patient sont parfois évaluées par différents observateurs, ce qui entraîne des discordances dans les dossiers cliniques. Enfin, dans un contexte de recherche, la variance dans l'estimation de l'âge osseux rend difficile la mise en place d'études multicentriques et la mise en commun des bases de données.

Toutes ces limitations démontrent le besoin d'avoir un outil qui assure une mesure fiable, reproductible et rapide pour la détermination de la maturité osseuse sur des radiographies du bassin. Un tel outil peut être développé grâce aux méthodes d'intelligence artificielle et apprentissage profond.

1.4 Apprentissage profond, vision par ordinateur et réseaux de neurones convolutifs.

L'apprentissage profond est une branche de l'intelligence artificielle dans laquelle une machine est capable de détecter des patrons et d'effectuer des tâches précises (ex. : classification, segmentation, prédiction) grâce à des exemples. Les données d'apprentissage sont étiquetées selon la classe à laquelle elles appartiennent, mais aucune autre information n'est inférée à la machine. Ainsi, l'apprentissage profond permet d'éviter que des humains ne spécifient expli-

citement les régions clés. Au lieu de cela, les caractéristiques les plus discriminantes et prédictives sont apprises à partir d'exemples étiquetés de façon hiérarchique, du plus abstrait au plus concret. Cette hiérarchie est d'ailleurs reflétée dans l'architecture en couche : les réseaux d'apprentissage profond sont une succession de modules simples où les représentations les plus abstraites sont calculées à la suite de termes moins abstraits. Enfin, la machine ajuste ses paramètres internes pour améliorer les prédictions en utilisant une méthode d'optimisation appelée rétropropagation de gradient (LeCun *et al.*, 2015). Récemment, des méthodes d'apprentissage profond ont été appliquées pour la segmentation, la détection et la classification. Les avancées majeures en termes d'apprentissage machine et de vision par ordinateur, la disponibilité de plus de puissance de calcul ainsi que la disponibilité de données d'apprentissage ont motivé la transition depuis les algorithmes conventionnels vers des approches d'apprentissage profond.

Les réseaux de neurones convolutifs sont inspirés du processus d'apprentissage humain et

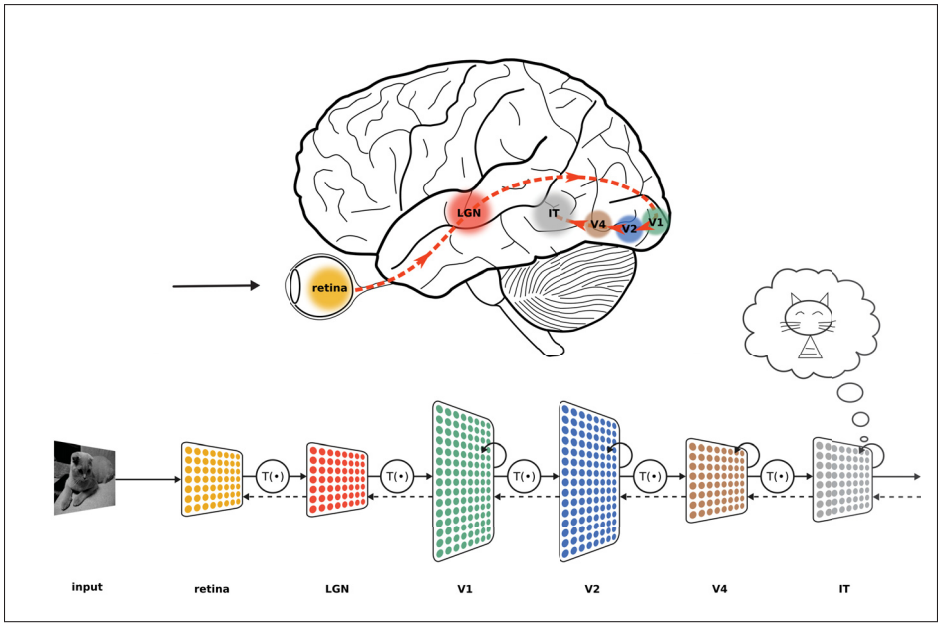

Figure 1.4 Analogie entre les procédés de vision humaine et les réseaux de neurones convolutifs reproduite avec la permission de (Kubilius, 2017)

des voies de reconnaissance visuelle, à travers lesquelles l'information est traitée de manière séquentielle avec une complexité croissante (figure : 1.4) : le cortex visuel primaire traite les

informations de bas niveau comme les bordures, la luminosité, qui correspondent au contexte. L'information est alors propagée à travers d'autres couches du cerveau, où les neurones sont plus actifs et réagissent de manière optimale aux stimuli complexes (par exemple les visages). Les neurones déterminent ainsi le "comment" en se basant sur les caractéristiques de haut niveau et la pondération apprise par l'expérience. En termes simples, le cerveau voit d'abord des contours, puis comprend qu'il s'agit d'un nez, une bouche, des yeux, et en se basant sur l'expérience prévoit avec une haute probabilité que ce qu'il perçoit soit en fait un visage.

De la même façon, les réseaux de neurones convolutifs commencent par extraire des caractéristiques brutes des images à l'aide de filtres, puis les informations sont propagées à travers différentes couches. Dans chaque couche, les neurones sont indépendants et connectés aux neurones de la couche suivante. Les neurones sont initialisés avec des pondérations et des biais affectés aléatoirement et optimisés avec l'entraînement. La sortie du réseau est un vecteur à N dimensions, ou le N correspond au nombre de classes.

1.4.1 Détermination de l'architecture optimale

"All models are wrong, but some are useful" George Edward Pelham Box.

Les réseaux de neurones convolutifs sont le type de réseau le plus adaptés à la tâche de classification. Il n'existe cependant pas de règles quant au choix de l'architecture en terme de profondeur, de paramètres, de poids ou d'outils d'optimisation. Il est ainsi judicieux de se baser sur la littérature pour effectuer un premier triage. Dans le domaine de l'imagerie médicale, de nombreuses études ont démontré l'utilité de réseaux entraînés sur la base de données ImageNet (Deng *et al.*) qui sont en fait un ensemble d'images naturelles annotées. L'état de l'art dans la classification d'images médicale a été atteint par les réseaux AlexNet, GoogLeNet, ResNet, Inception et VGG. Il était donc logique pour notre projet de commencer avec ces réseaux là. Après avoir déterminé notre objectif, qui est de performer au moins aussi bien que nos experts locaux et de minimiser l'amplitude de l'erreur de classification, Inception et VGG étaient nos meilleures options. Le choix final s'est porté sur VGG car la convergence s'est faite plus rapidement.

1.5 Automatisation de la détermination de la maturité osseuse

L'évaluation de la maturité osseuse fait partie intégrale de la pratique en orthopédie et en radiologie. Cependant, comme démontré au chapitre 1.3, les techniques actuelles sont laborieuses, chronophages et sujettes à des variabilités intra et inter-évaluateurs.

1.5.1 Techniques automatiques conventionnelles

Des outils pour l'estimation de l'âge osseux sur des radiographies de façon automatique existent depuis 1989. Le premier outil semi-automatique a été introduit par Nelson et Micheal, et utilisait des techniques de segmentation de carpogrammes (radiographies du poignet). D'autres outils ont continué à émerger, utilisant différentes techniques de vision par ordinateur comme les contours actifs (active shape models) (Niemeijer *et al.*, 2003), la segmentation, le filtrage et la détection de contours (Mansourvar *et al.*, 2013). Malgré les efforts mis en place pour automatiser la détermination de l'âge osseux, il a été fastidieux de voir un algorithme être utilisé en pratique. C'est en 2009 que BoneXpert, l'un des outils les plus populaires, a été développé puis commercialisé par Thodberg et Kreiborg et validé sur 4 ethnies différentes (Thodberg *et al.*, 2009). Cet outil est basé sur un algorithme en quatre étapes qui génère un modèle pour chaque os de la main et du poignet. Le modèle est ensuite comparé à un atlas de référence pour déterminer l'âge osseux. D'autres équipes ont également développé des algorithmes automatiques, toujours basés sur des régions d'intérêt manuellement spécifiées, comme Giordano et al. en 2010 qui ont basé leur algorithme sur la méthode Tanner Whitehouse (TW2), et qui détermine l'âge avec une précision de +/- 0.46 ans (Tanner *et al.*, 1975). Cependant, bien que ce type de méthode soit compréhensible, elles sont basées sur des régions d'intérêts prédéfinies par des humains, requièrent des images de haute qualité, et, puisqu'elles sont basées sur des règles heuristiques, risquent d'être moins efficaces dans les cas "limites".

1.5.2 Automatisation par apprentissage profond

L'utilisation de l'intelligence artificielle et de l'apprentissage machine pour la détermination de l'âge osseux a débuté en 1995. L'outil développé par Gross se basait sur l'extraction de caractéristiques discriminantes par un technicien, puis un système décisionnel était mis en place grace aux réseaux de neurones (Manzoor Mughal *et al.*, 2014). Il n'y avait cependant pas de différence entre l'utilisation de ce système et la classification manuelle basée sur la méthode de TW2. Avec l'amélioration des technologies et des puissances de calcul, une nouvelle vague de projets utilisant les réseaux de neurones convolutifs pour la détermination de l'âge osseux ont été développés, et ont démontré des résultats prometteurs. C'est le cas par exemple de l'équipe de Sampinato et al, qui ont en 2017 développé un réseau de neurones convolutif constitué de 5 couches. Leur réseau prend des radiographies de la main en entrée, et donne des prédictions d'âge osseux avec une précision dépassant l'état de l'art actuel (Spampinato *et al.*, 2017). Toutefois, le résultat le plus intéressant de leur étude concerne les régions que leur réseau a défini comme discriminantes. En effet, en visualisant l'apprentissage du réseau,il a été constaté que l'une des régions d'intérêt classiquement évaluées par la méthode TW inclut le troisième doigt (majeur). Cependant leur réseau a démontré que le troisième doigt n'est pas un critère discriminant, et a de plus pointé vers les phalanges du deuxième doigt (index), qui ne font normalement pas partie des régions d'intérêt dans la méthode de TW. Ce résultat est d'autant plus intéressant qu'il permet premièrement d'avoir une meilleure précision, un gain supérieur de temps et moins de variabilité, mais il démontre également un potentiel d'amélioration des connaissances médicales.

1.6 Conclusion de la revue de littérature

La revue de littérature présentée a dans un premier temps a permis de dresser des généralités sur l'anatomie de la colonne vertébrale et du des os du bassin, ainsi que d'introduire les notions d'âge osseux. Ces informations permettent de mieux comprendre la scoliose idiopathique adolescente, de clairement définir l'état de l'art dans l'imagerie pour la détermination de la ma-

turité osseuse au niveau du bassin tout en soulignant les aspects sur lesquels des améliorations peuvent être apportées.

Tout d'abord, il est d'un constat général que les méthodes de détermination de la maturité osseuse sont jusqu'à présent limitées aux radiographies de la main. À notre connaissance, aucune étude ne s'est intéressée à l'automatisation de la détermination de l'indice de Risser. La revue de littérature démontre donc la pertinence de notre étude en mettant ainsi en valeur son côté original ainsi que son intérêt clinique.

Parmi toutes les méthodes présentées, il est possible de distinguer deux principales catégories : a) les méthodes de détermination automatiques basées sur des méthodes de vision par ordinateur et des régions définies manuellement, et b) les méthodes basées sur des procédés d'intelligence artificielle, elles même séparées en méthodes semi-automatiques et apprentissage machine, et méthodes autonomes d'apprentissage profond.

De manière globale, les techniques utilisant l'apprentissage profond semblent obtenir des résultats plus satisfaisants, avec une réduction de la variabilité inter et intra observateur ainsi qu'une économie de temps et d'énergie. La revue de la littérature réalisée a donc motivé le choix de développer un outil d'apprentissage profond pour la détermination de l'indice de Risser. Il s'agit d'apporter une solution à un problème qui, couplé aux autres étapes de la prise en charge de la SIA, pourrait aider à la personnalisation du soin. Le choix d'utiliser les radiographies de face postéro-antérieures semble judicieux dans la mesure où cette technique est déjà la plus utilisée et permet une reproductibilité de cette étude.

CHAPITRE 2

DEEP LEARNING FOR AUTOMATIC RISSER STAGE ASSESSMENT.

Houda Kaddioui[1], Luc Duong[1], Julie Joncas[2], Christian Bellefleur[2], Imad Nahle[2], Olivier Chemaly[2], Marie-Lyne Nault[3], Stefan Parent[23], Guy Grimard[23], Hubert Labelle[23]

[1] Dept. of Software and IT Engineering, École de technologie supérieure, Montréal, Canada
[2] Division of Orthopedics, Centre Hospitalier Universitaire Sainte-Justine, Montréal, Canada
[3] Dept. of Surgery,Université de Montréal, Montréal, Canada

Paper submitted for publication in *Radiology : Artificial Intelligence*, October 2018

2.1 Introduction

Adolescent idiopathic scoliosis (AIS) is a 3D deformity of the spine, defined by a Cobb angle of at least 10 degrees. The prevalence of AIS is 1-3% and is more predominant in female patients (Hammond *et al.*, 2011). One study showed that AIS is a progressive disease causing significant health impacts like pain, musculoskeletal problems and psycho-sociological issues due to its occurrence in adolescence (Barton & Weinstein, 2018; Goldberg *et al.*, 1994; Mayo *et al.*, 1994). Management of AIS is mainly guided by the assessment of bone maturity since patients with significant growth potential have a greater risk of curve progression (Weinstein *et al.*, 2008). The Risser stage is the most commonly used indicator of bone maturity in AIS. In 1958, Risser introduced a comprehensive method observing the ossification of the iliac crest from conventional radiographs. Since then, two main classification systems emerged : the American and the French classifications. In the American classification, the ossification progression is divided in 6 stages, where stage 0 is a non-ossified iliac crest and 5 is a total fusion of the bones. The French classification divides the progression of the iliac ossification into thirds representing stages 0 to 3, stage 4 corresponds to the beginning of the crest's fusion to the iliac bone, and stage 5 is a complete fusion of the two bones (Hacquebord & Leopold, 2012). As of today, the Risser grade is widely accepted for assessing bone maturity and the progressive potential of AIS.

2.1.1 Raters variability in the assessment of the Risser stage.

Even with a clear clinical definition of the Risser stages, interpretation of plain radiographs is challenging due to : a) different image qualities between different acquisitions, b) variability between radiographic systems c) severe deformities where the strict frontal condition is no longer respected. Because of the rotated nature of the pelvis in AIS and subjective visual grading, an inter-observer and intra-observer variability was demonstrated and is accepted in the clinical practices. However, studies established a lack of consensus concerning the variability in the assessment of the Risser stage. Regarding the inter-observer variability, Goldberg et al. Goldberg *et al.* (1988) demonstrated a kappa of 0.80, and Dhar et al. (Dhar *et al.*, 1993) showed an agreement of 89.2%. In contrast, more recent studies showed a 50% agreement all stages combined, while Hammond and al, in agreement with a Shuren and al (Shuren *et al.*, 1992), showed moderate agreement between orthopedic surgeons and radiologists that can go up to three stages between the raters. Regarding the accuracy of the assessment, Izumi and al. suggest that anteroposterior radiographs reflect more on the iliac capping than the posteroanterior ones, leading to inaccuracies in the Risser stage assessment, while Reem et al. showed that it was an acceptable measure of the Risser stage (Hammond *et al.*, 2011; Izumi, 1995; Reem *et al.*, 2009; Sabour, 2018; Yang *et al.*, 2014). This evident variability can have a high impact when considering the therapeutic strategies and the outcomes. Moreover, patient's radiographs are sometimes graded by different observers, causing inconsistencies within the patient records. The controversy in the accuracy and reliability of the Risser grade can be resolved by a tool that guarantees a reliable and reproducible assessment without internal variability, and serves as a consistent second opinion. We propose such a computerized tool using deep learning methods(LeCun *et al.*, 2015).

2.1.2 Related works

Deep learning is a subset of artificial intelligence where a computer is able to detect patterns and make predictions leveraging example data. Deep learning avoids the need of having humans explicitly specifying key regions. Instead, the most predictive features are learned from

labelled examples as of a hierarchy of concepts, reflected in the architecture : deep learning networks are a stack of simple modules where more abstract representations are computed in terms of less abstract ones. Finally, the machine corrects its internal parameters to improve the predictions using an optimization method called back propagation(LeCun *et al.*, 2015; Abdolmanafi *et al.*, 2017). Recently, deep learning methods have been applied for segmentation, detection and classification. The major advances in machine learning and computer vision, and the availability of more computing power have motivated a shift from conventional algorithms towards deep learning approaches.

2.1.2.1 Deep learning for skeletal maturity.

Skeletal maturity evaluation is an integral part of the pediatric practice in general, and especially important for endocrinology, radiology and orthopedics. However, manual grading of a large number of radiographs is time consuming and getting a second opinion to reduce its variability is unfit for clinical settings. Previous studies have proposed automatic assessment of skeletal maturity, focusing mainly on carpograms. For instance, Thodberg and Kreiborg presented BoneXpert, a four-step algorithm generating bone models and comparing the output to a reference (Thodberg *et al.*, 2009). Such algorithms are highly comprehensive and easy to understand. Although useful and interpretable, they require high-quality images, and are based on heuristics which might fail to interpret borderline cases (Spampinato *et al.*, 2017). Deep learning has recently been introduced for radiographic assessment of skeletal maturity and have shown promising results. Spampinato et al. introduced an automatic bone age assessment on carpograms using a five layers convolutional neural network (Spampinato *et al.*, 2017). When looking at the key regions, the network suggested that some carpal regions accounted for by clinicians might not be relevant, while some new regions should be considered. The recent deep learning bone age assessment models yield satisfactory performance scores of 61% - 79% (Spampinato *et al.*, 2017; Lee *et al.*, 2017; Torres *et al.*, 2017). To the best of our knowledge, deep learning has not yet been applied for the assessment of the Risser stage on radiographs. Hence the goal of this study is to propose a new deep learning technique for the

automatic assessment of the Risser stage. We validate the performance of our method against observers by evaluating the intra and inter-observer variability.

2.2 Material and methods

This study was conducted at the orthopedic department of our institution. Institutional Review Board approval was obtained prior to the beginning of the study.

A total of 1830 posteroanterior EOS and standard radiographs were collected between 1999-2017 from the scoliosis clinic. Patients were aged 10 to 18 years old with a confirmed diagnosis of AIS. A female predominance in the dataset was 2.3 :1. Patients with bone abnormalities, irregular ossification, stunted growth and pelvis fractures were excluded. The Risser stage was collected from the patient's electronic records. The maximum Risser stage over the two iliac crests was set as the final label. This stage was evaluated by a trained technician and validated by an independent expert.The agreed upon stage was then used as the ground truth.

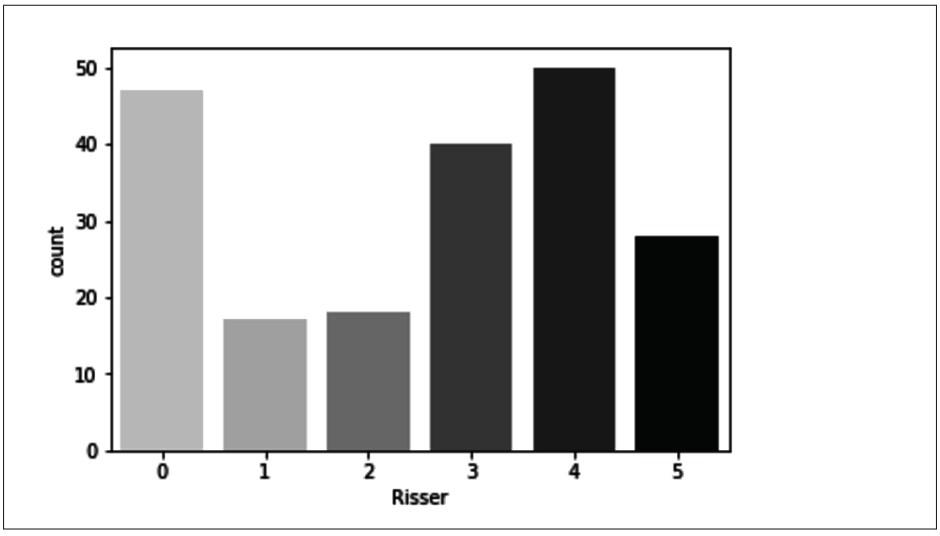

Figure 2.1 Distribution of the Risser grade in the radiographic database

2.2.1 Inter-observer and intra-observer agreement

To evaluate the intra and inter-observer variability, six graders were recruited. The group was composed of four orthopedic surgeons, one orthopedic fellow and one research nurse. All graders assess the Risser stage on a regular basis and are past the learning curve. A balanced sample of 200 shuffled radiographs was provided to each grader (figure 2.1). They were blinded about the sex, age, demographic information of the patients, the recorded Risser stage, and the assessment of their peers. Each grader classified the images independently, based on the American Risser classification. One grader classified the Risser stage on the same dataset, shuffled, one month later to evaluate intra-observer agreement.

Kappa coefficients (κ) measures the agreement between graders while accounting for the effect of chance. If the graders are in complete agreement, κ=1, while if there is no agreement κ=0. Since the group has more than two graders, the Fleiss variation was used (Fleiss & Cohen, 1973).

Tableau 2.1 Pairwise Kappa value of the observers (Obs), the ground truth and the proposed automatic method

Observers	**Obs 1**	**Obs 2**	**Obs 3**	**Obs 4**	**Obs 5**	**Obs 6**	**Automatic grading method**	**Ground truth**
Obs 1	1.00	0.62	0.53	0.55	0.71	0.68	0.64	0.63
Obs 2	-	1.00	0.50	0.50	0.62	0.55	0.54	0.57
Obs 3	-	-	1.00	0.59	0.57	0.65	0.58	0.52
Obs 4	-	-	-	1.00	0.53	0.58	0.57	0.49
Obs 5	-	-	-	-	1.00	0.66	0.69	0.60
Obs 6	-	-	-	-	-	1.00	0.60	0.52
Automatic grading method	-	-	-	-	-	-	1.00	0.72
Ground truth	-	-	-	-	-	-	-	1.00

Tableau 2.2 Pairwise percentage of agreement for the observers (Obs), the ground truth and the proposed automatic method

Observers	**Obs 1**	**Obs 2**	**Obs 3**	**Obs 4**	**Obs 5**	**Obs 6**	**Automatic grading method**	**Ground truth**
Obs 1	100.0	71.0	68.5	64.5	81.0	75.5	72.0	71.0
Obs 2	-	100.0	62.5	61.0	71.0	65.5	65.0	66.0
Obs 3	-	-	100.0	68.0	68.5	74.5	68.5	62.5
Obs 4	-	-	-	100.0	63.5	67.5	66.4	59.0
Obs 5	-	-	-	-	100.0	74.5	76.0	69.0
Obs 6	-	-	-	-	-	100.0	67.5	62.0
Automatic grading method	-	-	-	-	-	-	100.0	78.0
Ground truth	-	-	-	-	-	-	-	100.0

The graders were organized in two groups : senior experts (more than twenty years of experience) and new experts (less than ten years of experience). The overall agreement was first computed, followed by the agreement within groups. The results were compared to Landis and Koch's agreement scale : lower than zero corresponds *to less than chance agreement*, 0.01–0.20 *slight agreement*, 0.21– 0.40 *fair agreement*, 0.41–0.60 *moderate agreement*, 0.61–0.80 *substantial agreement* and 0.81–0.99 *almost perfect agreement* (Landis & Koch, 1977). Groupwise and pairwise percentages of agreement were computed for a better interpretation of the observers' agreement.

2.2.2 Automatic Risser grading

Convolutional neural networks (CNNs) are a subtype of deep learning. The architecture of CNNs is inspired by the human hierarchical learning process and visual recognition pathways where information is processed sequentially with increased complexity (LeCun *et al.*, 2015; Blumberg & Kreiman, 2010) : primary visual cortex processes low-level information (edges, luminosity etc.) to determine the how/where. The information is then propagated through other layers of the brain, where neurons along the ventral stream are more active and respond optimally to complex stimuli (e.g. faces), finding the what based on high-level features and weights learned from experience (Poggio & Bizzi, 2004). Similarly, CNNs start by retrieving rough features from images using filters, and the information is propagated through different layers. In each layer, the neurons are independent and connected to neurons in the next layer. The neurons are inferred with weights and biases initialized randomly and optimized with training.

Training deep learning networks requires a large number of annotated images. Since the number of radiographs was limited, we applied transfer learning. This approach consisted of reusing a CNN trained on a large dataset (e.g. natural images) and adjusting its parameter to better fit our dataset. Transfer learning has been proven effective in practice for medical imaging (Abdolmanafi *et al.*, 2017; Litjens *et al.*, 2017).

A simple preprocessing of the radiographs was done : The images were first cropped along the smallest edge then resized, keeping the aspect size ratio while including the entire pelvis. A median filter was applied afterwards to remove the salt and pepper noise. The dataset was then split into training and validation set at an 80% - 20% ratio. A third subset was left as a second testing set used for the validation of the accuracy against the experts as mentioned above. When the images were input to the network, convolution filters of fixed size created a feature map by sliding over the entire image following a fixed stride. Convolution layers were followed by rectified linear unit (ReLU) layer, adding non-linearity and making the network generalizing well to any type of function. Afterwards, a pooling layer was used to sample over the output of the previous layer, only keeping the most valuable information by retaining the maximum value in a given N*N window. The final layers of the network were specifically developed to train on the Risser grading task. This new set was randomly initialized and connected to the body of the original network. The fully connected layers result in a computed output of size 1*1*C where C is the number Risser stages (figure 2.2).

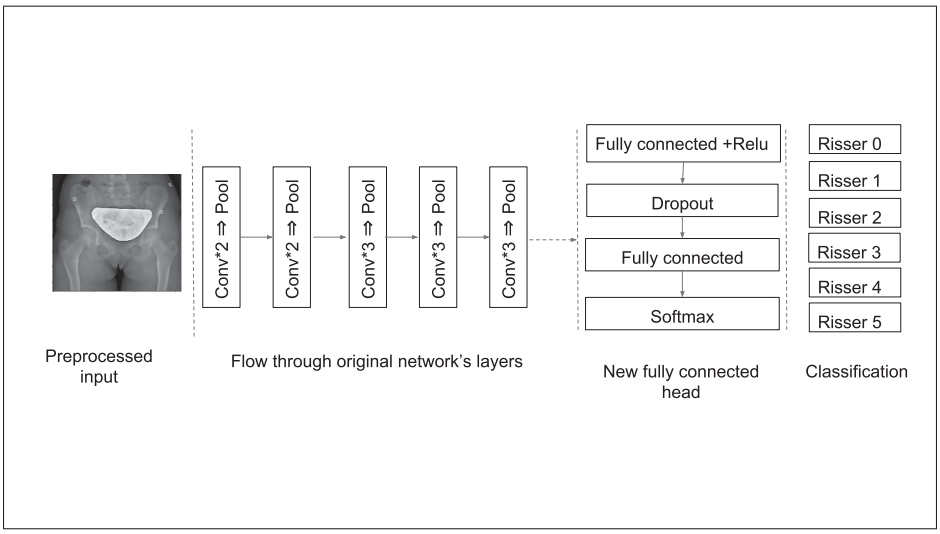

Figure 2.2 Feature extraction and classification workflow with convolutional neural networks. The output of the proposed method is the Risser grade (0-5)

The model parameters were initialized to pretrained weights optimized for the ImageNet dataset. To keep the parameters of the trained model, the first step was to freeze the superficial layers and only train the new layers over multiple iterations. This avoids a propagation of the gradient over the entire network and prevents losing the discriminating parameters for the kernels, while allowing the filters to learn new parameters. After 30 iterations, the layers were "unfrozen", and training continued until sufficient accuracy was obtained, with a very low learning rate. Stochastic gradient descent was used as an optimization algorithm to correct the predictions and guide the network toward accurate weights. After determining the final parameters, the training was done for 10 folds to control for the effect of chance. To compare our network to the observers' performance, we computed the accuracy against the agreement interval of the different grader groups. The software was developed in Python 2.7 using the Keras library with the Tensorflow library for deep learning (Abadi *et al.*, 2016)

2.3 Results

2.3.1 Inter and intra-observer agreement

Overall agreement between observers was fair. Senior experts (Obs. 5 and Obs. 6) had a kappa coefficient of 0.6 and an agreement of 74.5%. New experts (Obs. 1-4) had a kappa coefficient of 0.58, with an agreement of 41.5%. The pairwise kappa coefficients and percentage of agreement for all observers are presented in (table 2.1) and (table 2.2). The pairwise agreement ranged from fair to moderate. Intra-observer agreement was substantial (κ=0.62) and is lower than the inter-observer variability. The inter-class coefficient was computed to analyze the percentage of agreement of the experts with the ground truth (true Risser stage). Figure 2.3 shows the performances of each expert and the group performance over each class. Analyzing the confusion matrix revealed high performances on Risser stage 0, 1 and 5 while stages 3 and 4 had the most variability.

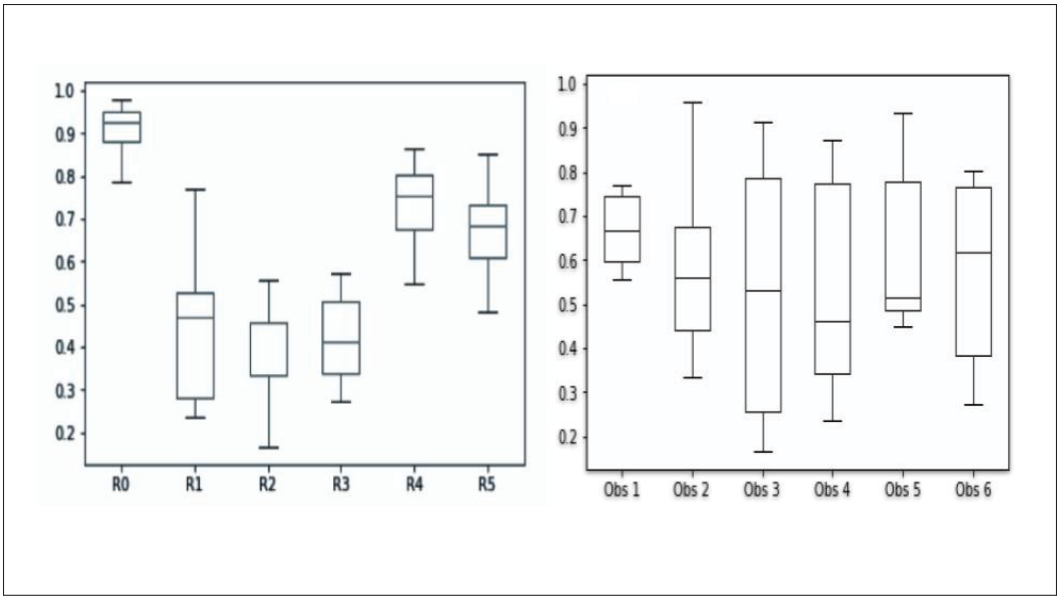

Figure 2.3 Performance of each grader in grading the test set (left) and performance of all the graders for each Risser stage(right). Obs= Obsever; R=Risser

2.3.2 Automatic Risser grading method

Our model was tested on the same dataset given to the graders group. The automatic grading method showed a substantial agreement with the ground truth (κ=0.72), and an overall accuracy of 78% with very low variability (standard variation = 2.30). Analyzing the output of the network showed a variability of classification limited to two stages, while the graders could have a variability of three or more stages (Figure 2.4).

Moreover, the misclassified images correspond to the most controversial images with the less agreement between the observers (Figure 2.5). The computing time was less than one second per image. The training phase, performed only once, took eight hours on a professional workstation with high-end graphical processing unit (NVIDIA Titan X).

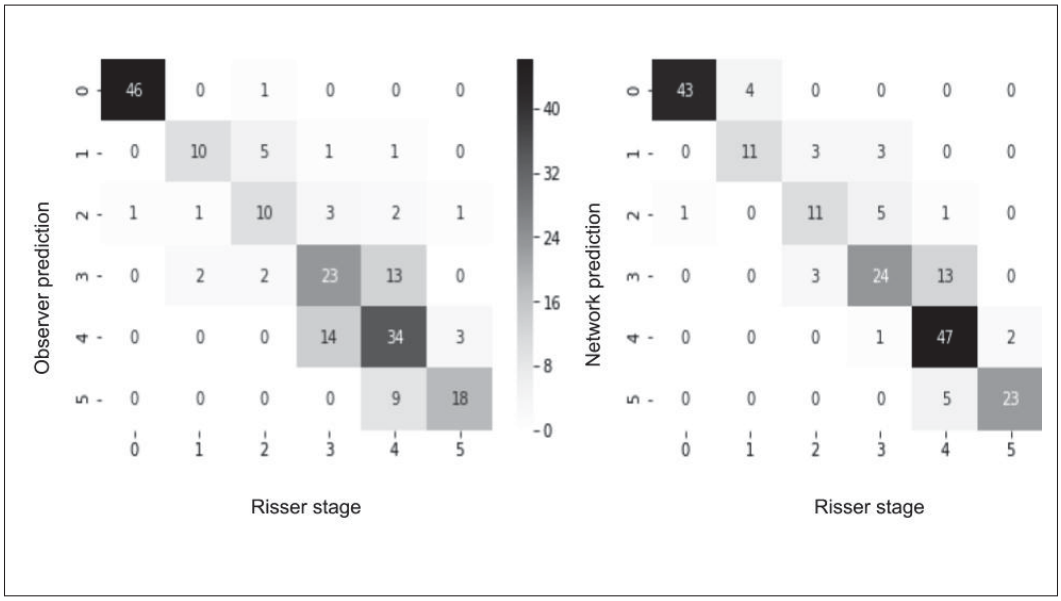

Figure 2.4 Confusion matrix for one of the observers (left) and the automatic grading method (right). This matrix illustrate correctly and incorrectly classified samples. The rows of the matrix show the values indicated by the observer while the columns show the ground truth. The values on the diagonal of the matrix illustrate the number of samples correctly classified by Risser grade. The values above and below each value of the diagonal show misclassified samples

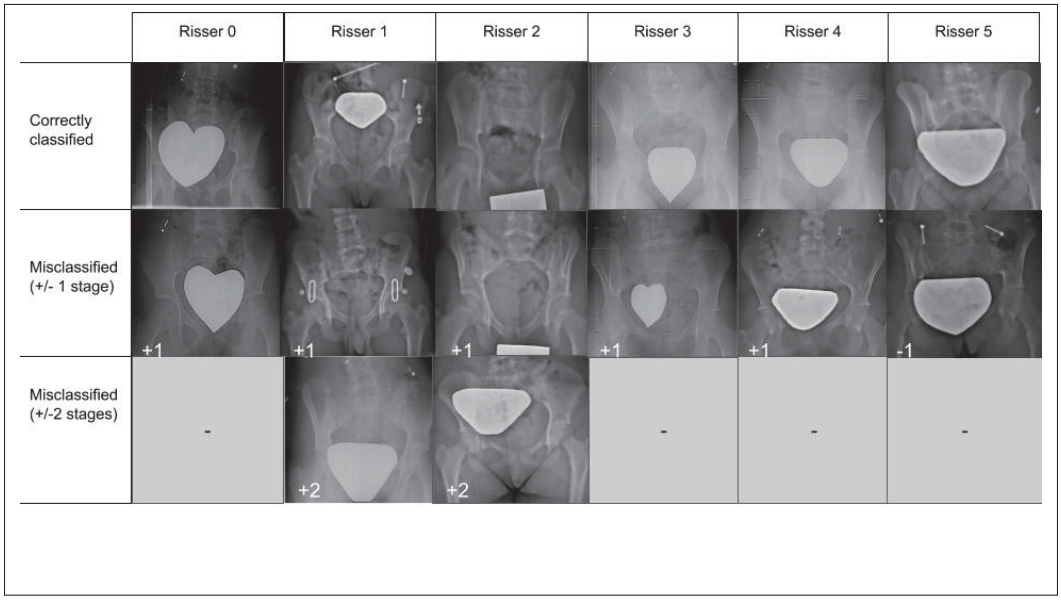

Figure 2.5 Sample radiographic images correctly classified by automatic grading method (top) and misclassified by one grade (2nd row) and two grades (3rd row)

2.4 Discussion

The Risser stage is a widely used indicator of skeletal maturity and progression potential of AIS. Although this stage is comprehensive and easy to implement, several authors raised concerns regarding its validity and reliability. Studies suggest that the Risser system is subject to inter-observer variability, is not a good reflection of the velocity of the curve progression, and lacks responsivness (because of a low sensitivity to rapid acceleration phases) (Reem *et al.*, 2009). Sanders and al. introduced a new classification of bone maturity based on wrist radiographs (Sabour, 2018). A study comparing the Risser and Sanders classifications showed a higher kappa coefficient for the latter (Minkara *et al.*, 2018). Following this theory, Nault et al. proposed a new Risser classification that includes the triradiate cartilage(Nault *et al.*, 2010). Also, Hresko and al. proposed a revised classification with eight Risser stages, combining the American and French classifications with the triradiate cartilage ossification. Their inter-observer evaluation produced insufficient agreement (Hresko *et al.*, 2018). All these studies show a common concern regarding the grading variability among experts.

Previous literature reports show a kappa value of 0.31 - 0.80. This broad range underlines the need for normalized databases, intra and inter-observer studies, and for developing automated grading systems. Our results found a fair to moderate agreement, matching the literature's highest agreement values. However, the interpretation of kappa values must consider two factors : first, the null hypothesis in a medical context should not be set as κ=0, but a minimal acceptable agreement should be decided upon. To our knowledge, no such value was defined, hence the need to obtain the best possible agreement. The second aspect is the effect of variability on the therapeutic decision : a study showed that the variability in assessing the Risser stage led to 21% over-referral rate (Hammond *et al.*, 2011). In the clinical context, this means additional medical costs, missing classes, and radiation exposures, added to the impact of the treatment that can be overwhelming for adolescents (Goldberg *et al.*, 1994; Weinstein *et al.*, 2008). Getting a second opinion might reduce this variability and thus reduce the propagation of an error bias within the patient's files as well as the rate of overdiagnosis and over referrals. However, a second opinion is usually not easily available. Since our network have been trained

on an agreement of two experts and validated on a group of six other graders, its classification comes as a second opinion. Moreover, some of the factors like the time, physical state or work load can reduce the accuracy of the classification. However, the network is invariant and independent of these factors.

This is the first study to use deep learning technique for Risser grading. Our results illustrate that a CNNs can be used to assign the Risser grade with satisfying accuracy. An automatic method is appealing since computerized approaches are highly predictive and give consistent output for the same input without internal variability. Furthermore, the result is given within seconds, which is an interesting feature for radiologists and orthopedic surgeons. Finally, the network was trained to learn the most specific and invariant features, making it robust against different image variations, rotations and contrasts thus overcoming the limitations of the Risser grading system. Although different authors question the reliability of the Risser stage, this study is promising and shows the potential for a more accurate detection on radiographs. There are some limitations to this work : the ground truth was used based on the agreement of two observers, meaning that the network could be less accurate on a noisier dataset. Our work can be improved by collecting more radiographs and having additional graders agree on the final label. Finally, since the network was only trained on healthy patient's radiographs, an improvement could be made by including patients with irregular ossification, growth latency and other bone disorders. While we achieved 78% accuracy, additional performance and reliability gain could be reached by diversifying the dataset.

2.4.1 Conclusion

An automatic Risser grading method was developed using a convolutional neural network, a deep learning approach. In addition, we evaluated the intra and inter-observer variability at our institution. Our automatic method was able to perform within the known inter-observer variability, without internal variability. These results pave the road for more investigation on the feasibility of integrating automatic radiographic methods in clinical settings and its usefulness for the management of AIS.

CHAPITRE 3

DISCUSSION

Cette étude consistait à développer un outil automatisant la classification de Risser en utilisant des données d'imagerie à rayon-X et des méthodes d'intelligence artificielle. Nous avons ensuite validé nos résultats en comparant la performance de notre outil à la variabilité inter-observateur d'experts sélectionnés au sein de notre institution. Dans cette section, nous allons discuter les résultats obtenus ainsi que les défis rencontrés pendant la conception expérimentale, les compromis entre les choix effectués lors de la mise en œuvre méthodologique et les limitations résultant de ces choix de conception. Enfin, nous évoquerons les perspectives d'avenir et terminerons avec une réflexion sur les potentiels thèmes de recherche future.

Pour accomplir notre projet, nous avons répondu aux sous objectifs suivants : 1) Collecte et préparation des données 2) Développement et entraînement de l'algorithme d'apprentissage profond 3) Détermination de la variabilité inter et intra-observateur locale et évaluation des performances du détecteur automatique. Pour atteindre chacun des sous objectifs de ce projet, nous avons fait face à plusieures limitations en raison de multiples facteurs, notamment la nature même des données médicales, de la subjectivité dans les pratiques cliniques, et l'hétérogénéité dans la population cible, ainsi que les limites intrinsèques des algorithmes d'apprentissage machine et d'apprentissage profond. Ces limitations sont discutées ci-dessous ainsi que les compromis et choix faits pendant le processus de conception expérimentale.

3.1 Collecte et préparation des données

La première étape dans notre processus était d'assembler un ensemble de données suffisamment large pour pouvoir entraîner un réseau de neurones artificiels. En effet, une des limitations intrinsèques des réseaux de neurones convolutifs est que l'échantillon d'exemples doit être assez grand pour apprendre des caractéristiques discriminantes. Le second choix à faire était la sélection de la population à étudier : pour pouvoir tester notre hypothèse, nous avons décidé de ne sélectionner que les patients n'ayant pas de comorbidités, et plus spécifiquement pas

de troubles du développement osseux autre que la SIA. Ceci rend notre échantillon moins représentatif de la population générale, mais permet d'avoir des données plus homogènes. Une amélioration logique serait donc d'inclure les patients qui faisaient partie de nos critères d'exclusion. Concernant le pré-traitement des images, nous avons décidé de limiter notre intervention à la réduction de la taille en préservant le ratio pour ne pas modifier l'anatomie. La seconde partie du pré-traitement consistait à retirer le bruit "poivre et sel" à l'aide d'un filtre médian. Nous avons choisi de ne pas appliquer de méthodes de normalisation, car ceci rendrait plus laborieuse l'inclusion de nouveau cas dans le cas d'un entraînement dynamique : en effet, les techniques de normalisation suggèrent de considérer l'ensemble des données, et donc répéter la tâche à chaque fois qu'une nouvelle image est ajoutée à l'ensemble de données. De plus, un excès de pré-traitement réduit le potentiel de généralisation de l'algorithme.

3.2 Entraînement du réseau de neurones convolutif

Pour accomplir ce sous objectif, nous avons choisi d'utiliser un réseau de neurones entraînés sur des images d'un domaine différent et de l'ajuster aux données médicale, puis de valider ses performances en le comparant à nos experts. Une comparaison directe de notre méthodologie et nos résultats à la littérature est difficile car il s'agit de la première étude automatisant l'assignation du stade de Risser. Cependant, le méthodologie suivie est couramment retrouvée dans les études appliquant l'apprentissage profond à l'imagerie médicale. Une limitation du processus d'entraînement est que les images ont uniquement était annotées par deux experts. La précision des annotations peut être améliorée en mobilisant un plus grand groupe pour réduire la variabilité.

3.3 Évaluation de la variabilité inter et intra-évaluateur.

Bien que le stade de Risser soit compréhensible, plusieurs auteurs ont exprimé des préoccupations concernant sa validité et sa fiabilité. Des études suggèrent que la classification est sujette à une variabilité inter-observateur, ne reflète pas bien la vitesse de progression de la courbe et n'est pas sensible aux périodes d'accélération de la croissance. Sanders et al. ont ainsi introduit

une nouvelle classification de la maturité osseuse basée sur les radiographies du poignet. Une étude comparant les classifications de Risser et de Sanders a montré un coefficient kappa plus élevé pour ce dernier (Sabour, 2018). En se basant sur cette théorie, Nault et al. ont proposé une nouvelle classification de Risser qui inclut le cartilage triradié (Nault *et al.*, 2010). Aussi, Hresko et al. ont proposé une classification révisée avec huit stades de Risser, combinant les classifications américaine et française avec l'ossification du cartilage triradié(Hresko *et al.*, 2018). Leur évaluation de la variabilité inter-observateurs a démontré un accord insuffisant pour que cette stadification soit implémentée. Toutes ces études montrent une préoccupation commune concernant la variabilité dans l'estimation de l'âge osseux parmi les experts. Des rapports de littérature antérieurs montrent une valeur kappa de 0,31 à 0,80. Ce large intervalle souligne le besoin accru de bases de données normalisées, d'études intra et inter-observateurs et de développement de systèmes de classement automatisés.

Nos résultats ont trouvé un accord juste à modéré, correspondant aux valeurs d'accord les plus élevées de la littérature. Cependant, l'interprétation des valeurs kappa doit tenir compte de deux facteurs : premièrement, l'hypothèse nulle dans un contexte médical ne devrait pas être définie par k = 0, mais il faudrait convenir d'un accord minimal acceptable. À notre connaissance, aucune valeur de ce type n'a été définie, d'où la nécessité d'obtenir le meilleur accord possible. Le deuxième aspect est l'effet de la variabilité sur la décision thérapeutique : une étude a montré que la variabilité dans l'évaluation du stade Risser entraînait un taux de surréférence de 21%. Dans le contexte clinique, cela signifie des coûts médicaux supplémentaires, une augmentation de l'absentéisme scolaire et plus d'expositions aux radiations s'ajoutant à l'impact du traitement qui peut être lourd pour les adolescents.

CONCLUSION ET RECOMMANDATIONS

L'objectif de la recherche était de développer un outil automatisant la classification de Risser grâce à un réseau de neurones convolutif. Notre recherche a permis d'apporter des contributions intéressantes : tout d'abord, il s'agit de la première tentative d'automatiser le processus, avec des résultats que nous jugeons très prometteurs. La principale contribution de ce travail est de mettre à disposition des professionnels de la santé un outil fiable, dont les résultats sont répétables et indépendants de facteurs externes. Cet outil peut être utilisé au delà du contexte de SIA, dans toutes les situations ou l'information sur l'âge biologique est nécessaire, comme la prédiction de la taille maximale, l'estimation de la croissance à venir, ou pour préparer des interventions de remodelage osseux. Il permettra de réduire la variabilité dans l'estimation de l'indice de Risser et ainsi diminuer la propagation d'un biais d'erreur dans les dossiers médicaux. Enfin, cet outil est facile à implémenter, financièrement accessible dans le contexte hospitalier et produit des résultats en quelques secondes. La validation de la méthode proposée a été réalisée sur un ensemble de données exclusif permettant de comparer les performances de notre détecteur automatique à celle de nos experts. Notre outil performait au-dessus de l'intervalle de variabilité de nos experts.

Bien que les résultats soient encourageants pour l'utilisation d'une méthode d'apprentissage profond, certaines limites sont soulevées par cette étude. Tout d'abord, notre outil a été développé sur des données de patients atteints de SIA, mais en excluant les autres anomalies osseuses. Cette approche peut être critiquée car la SIA peut survenir sur un terrain avec des comorbidités et il est fréquent de retrouver des cas d'ossification irrégulière. Une des principales améliorations serait donc de pouvoir entraîner un second réseau de neurones artificiels à détecter de tels cas. Une seconde amélioration de la performance pourrait être atteinte en utilisant des radiographies de face et de profil, permettant de trianguler l'information et de surpasser les limites de la radiographie de face "stricte". Enfin, il serait judicieux d'augmenter le nombre d'experts qui annotent les images pour aboutir à un outil plus robuste.

3.4 Perspectives d'avenir

L'utilisation de l'apprentissage profond dans le domaine médical est de plus en plus fréquente. Nous pensons que notre outil pourrait être amélioré en le rendant dynamique, c'est à dire qu'il serait de mieux en mieux entraîné avec le nombre de données qui seraient ajoutées. Un outil dynamique pourrait alors être ajouté à l'arsenal diagnostique mis à disposition des cliniciens. Une perspective serait de pouvoir intégrer des données supplémentaires de patients comme le stade de Sanders ou les informations démographiques pour aboutir à un outil plus complet qui assisterait les professionnels de la santé dans leur décision. Avec la disponibilité des ressources en terme de puissance de calcul, les résultats pourraient être donnés en temps réel, pourraient devenir plus accessible et seraient un nouveau pas envers une médecine personnalisée. Un tel outil pourrait être très utile dans le cadre de télé-médecine, dans les régions où l'accès au soin est limité et où il existe un manque crucial de spécialistes, comme c'est le cas dans les pays pauvres et les pays en voie de développement. Enfin, les réseaux de neurones artificiels peuvent avoir leur place dans l'éducation continue des professionnels de santé : en effet comme reporté précédement, il serait bon de prendre en compte les informations obtenues par un réseau pour réajuster les connaissances médicales. Les réseaux entraînés sur des consensus d'un groupe de professionnels peuvent également être une référence par rapport aux performances individuelles : un professionnel qui n'est pas en accord avec le réseau n'est théoriquement pas en accord avec le groupe de professionnels. Cette information peut permettre de proposer une formation complémentaire et éviter les erreurs de diagnostic.

L'avenir de l'intelligence artificielle dans le domaine de l'imagerie médicale et de la santé en général semble prometteur. Une utilisation réfléchie de ces outils, prenant en compte l'objectif initial qui est le bien être du patient, et dans un cadre éthique, pourrait radicalement améliorer la pratique médicale actuelle et l'accès au soin vers une médecine pro-active et collaborative.

BIBLIOGRAPHIE

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G. & Isard, M. (2016). TensorFlow : A System for Large-Scale Machine Learning. *OSDI*, 16, 265–283.

Abdolmanafi, A., Duong, L., Dahdah, N. & Cheriet, F. (2017). Deep feature learning for automatic tissue classification of coronary artery using optical coherence tomography. *Biomed. opt. express, boe*, 8(2), 1203–1220. doi : 10.1364/BOE.8.001203.

Barton, C. B. & Weinstein, S. L. (2018). Adolescent Idiopathic Scoliosis : Natural History. Dans Machida, M., Weinstein, S. L. & Dubousset, J. (Éds.), *Pathogenesis of Idiopathic Scoliosis* (pp. 27–50). Tokyo : Springer Japan. doi : 10.1007/978-4-431-56541-3_2.

Blumberg, J. & Kreiman, G. (2010). How cortical neurons help us see : visual recognition in the human brain. *J clin invest*, 120(9), 3054–3063. doi : 10.1172/JCI42161.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. Imagenet : A large-scale hierarchical image database. 2.

Dhar, S., Dangerfield, P. H., Dorgan, J. C. & Klenerman, L. (1993). Correlation between bone age and Risser's sign in adolescent idiopathic scoliosis. *Spine*, 18(1), 14–19.

Fleiss, J. L. & Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and psychological measurement*, 33(3), 613–619. doi : 10.1177/001316447303300309.

Goldberg, M. S., Mayo, N. E., Poitras, B., Scott, S. & Hanley, J. (1994). The Ste-Justine Adolescent Idiopathic Scoliosis Cohort Study. Part II : Perception of health, self and body image, and participation in physical activities. *Spine*, 19(14), 1562–1572.

Goldberg, M. S., Poitras, B., Mayo, N. E., Labelle, H., Bourassa, R. & Cloutier, R. (1988). Observer Variation in Assessing Spinal Curvature and Skeletal Development in Adolescent Idiopathic Scoliosis. *Spine*, 13(12), 1371–1377. Repéré à https://insights.ovid.com/crossref?an=00007632-198812000-00008.

Gorman, K. F., Julien, C. & Moreau, A. (2012). The genetic epidemiology of idiopathic scoliosis. *European spine journal*, 21(10), 1905–1919.

Hacquebord, J. H. & Leopold, S. S. (2012). In Brief : The Risser Classification : A Classic Tool for the Clinician Treating Adolescent Idiopathic Scoliosis. *Clin orthop relat res*, 470(8), 2335–2338. doi : 10.1007/s11999-012-2371-y.

Hammond, K. E., Dierckman, B. D., Burnworth, L., Meehan, P. L. & Oswald, T. S. (2011). Inter-observer and intra-observer reliability of the Risser sign in a metropolitan scoliosis screening program. *Journal of pediatric orthopaedics*, 31(8), e80–e84. Repéré à http://journals.lww.com/pedorthopaedics/Abstract/2011/12000/Inter_Observer_and_Intra_Observer_Reliability_of.12.aspx.

Herman, R., Mixon, J., Fisher, A., Maulucci, R. & Stuyck, J. (1985). Idiopathic scoliosis and the central nervous system : a motor control problem. the harrington lecture, 1983. scoliosis research society. *Spine*, 10(1), 1–14.

Horne, J. P., Flannery, R. & Usman, S. (2014). Adolescent idiopathic scoliosis : Diagnosis and management. *American family physician*, 89(3), 193–198.

Hresko, M. T., Troy, M., Miller, P., Price, N., Talwalkar, V., Zaina, F., Donzelli, S. & Negrini, S. (2018). Risser plus sign : a new grading system to classify skeletal maturity in scoliosis patients. *European spine journal*. doi : https://doi.org/10.1007/s00586-018-5821-8.

Izumi, Y. (1995). The accuracy of Risser staging. *Spine*, 20(17), 1868–1871.

Kubilius, J. (2017). Ventral visual stream. figshare. doi : 10.6084/m9.figshare.106794.v3.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. doi : 10.1038/nature14539.

Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K., Choy, G. & Do, S. (2017). Fully Automated Deep Learning System for Bone Age Assessment. *J digit imaging*, 30(4), 427–441. doi : 10.1007/s10278-017-9955-8.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B. & Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. *arxiv :1702.05747 [cs]*. Repéré à http://arxiv.org/abs/1702.05747. arXiv : 1702.05747.

Mansourvar, M., Ismail, M. A., Herawan, T., Gopal Raj, R., Abdul Kareem, S. & Nasaruddin, F. H. (2013). Automated Bone Age Assessment : Motivation, Taxonomies, and Challenges. *Computational and mathematical methods in medicine*, 2013, 1–11. doi : 10.1155/2013/391626.

Manzoor Mughal, A., Hassan, N. & Ahmed, A. (2014). Bone Age Assessment Methods : A Critical Review. *Pak j med sci*, 30(1), 211–215. doi : 10.12669/pjms.301.4295.

Marshall, W. A. & Tanner, J. M. (1969). Variations in pattern of pubertal changes in girls. *Archives of disease in childhood*, 44(235), 291–303.

Marshall, W. A. & Tanner, J. M. (1970). Variations in the pattern of pubertal changes in boys. *Archives of disease in childhood*, 45(239), 13–23.

Mayo, N. E., Goldberg, M. S., Poitras, B., Scott, S. & Hanley, J. (1994). The Ste-Justine Adolescent Idiopathic Scoliosis Cohort Study. Part III : Back pain. *Spine*, 19(14), 1573–1581.

Minkara, A., Bainton, N., Tanaka, M., Kung, J., DeAllie, C., Khaleel, A., Matsumoto, H., Vitale, M. & Roye, B. (2018). High Risk of Mismatch Between Sanders and Risser Staging in Adolescent Idiopathic Scoliosis : Are We Guiding Treatment Using the Wrong Classification ? *Journal of pediatric orthopaedics*, 1. doi : 10.1097/BPO.0000000000001135.

Nault, M.-L., Parent, S., Phan, P., Roy-Beaudry, M., Labelle, H. & Rivard, M. (2010). A Modified Risser Grading System Predicts the Curve Acceleration Phase of Female Adolescent Idiopathic Scoliosis :. *The journal of bone and joint surgery-american volume*, 92(5), 1073–1081. doi : 10.2106/JBJS.H.01759.

Niemeijer, M., van Ginneken, B., Maas, C. A., Beek, F. J. A. & Viergever, M. A. (2003). Assessing the skeletal age from a hand radiograph : automating the Tanner-Whitehouse method. pp. 1197. doi : 10.1117/12.480163.

Phan, P., Ouellet, J., Mezghani, N., De Guise, J. A. & Labelle, H. (2015). A rule-based algorithm can output valid surgical strategies in the treatment of ais. *European spine journal*, 24(7), 1370–1381.

Poggio, T. & Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, 431(7010), 768–774. doi : 10.1038/nature03014.

Ramirez, L., Durdle, N. G., Raso, V. J. & Hill, D. L. (2006). A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topography. *Ieee transactions on information technology in biomedicine*, 10(1), 84–91.

Reem, J., Carney, J., Stanley, M. & Cassidy, J. (2009). Risser sign inter-rater and intra-rater agreement : is the Risser sign reliable ? *Skeletal radiology*, 38(4), 371–375. doi : 10.1007/s00256-008-0603-8.

Sabour, S. (2018). Reliability of the Sanders Classification Versus the Risser Stage ; Avoid Misinterpretation. *Journal of pediatric orthopaedics*, 38(1), e29. doi : 10.1097/BPO.0000000000001059.

Sardjono, T. A., Wilkinson, M. H., Veldhuizen, A. G., van Ooijen, P. M., Purnama, K. E. & Verkerke, G. J. (2013). Automatic cobb angle determination from radiographic images. *Spine*, 38(20), E1256–E1262.

Shuren, N., Kasser, J. R., Emans, J. B. & Rand, F. (1992). Reevaluation of the use of the Risser sign in idiopathic scoliosis. *Spine*, 17(3), 359–361.

Simony, A., Carreon, L. Y., H ?jmark, K., Kyvik, K. O. & Andersen, M. O. (2016). Concordance rates of adolescent idiopathic scoliosis in a danish twin population. *Spine*, 41(19), 1503. doi : 10.1097/BRS.0000000000001681.

Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M. & Leonardi, R. (2017). Deep learning for automated skeletal bone age assessment in X-ray images. *Medical image analysis*, 36, 41–51. doi : 10.1016/j.media.2016.10.010.

Tanner, J. M., Whitehouse, R. H., Marshall, W. A. & Carter, B. S. (1975). Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height. *Arch. dis. child.*, 50(1), 14–26.

Thodberg, H. H., Kreiborg, S., Juul, A. & Pedersen, K. D. (2009). The BoneXpert method for automated determination of skeletal maturity. *Ieee trans med imaging*, 28(1), 52–66. doi : 10.1109/TMI.2008.926067.

Torres, F., Bravo, M. A., Salinas, E., Triana, G. & Arbeláez, P. (2017). Bone age detection via carpogram analysis using convolutional neural networks. 10572, 1057217. doi : 10.1117/12.2285949.

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

Yang, J. H., Bhandarkar, A. W., Suh, S. W., Hong, J. Y., Hwang, J. H. & Ham, C. H. (2014). Evaluation of accuracy of plain radiography in determining the Risser stage and identification of common sources of errors. *Journal of orthopaedic surgery and research*, 9(1), 101.