UNIVERSITÉ DE LILLE

FACULTÉ DE MÉDECINE HENRI WAREMBOURG

Année : 2021

THÈSE POUR LE DIPLÔME D'ÉTAT

DE DOCTEUR EN MÉDECINE

Développement et évaluation de modèles d'intelligence artificielle pour la détection des lésions post-traumatiques du coude de l'enfant en radiographie.

Présentée et soutenue publiquement le 1er Juillet 2021 à 16h
au Pôle Formation
par Clémence ROZWAG

_____

JURY

Présidente :
Madame le Professeur Anne COTTEN

Assesseurs :
Monsieur le Professeur Xavier DEMONDION
Monsieur le Professeur Philippe PREUX

Directeur de thèse :
Monsieur le Docteur Thibaut JACQUES

_____

AVERTISSEMENT

La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

TABLE DES MATIÈRES

ABRÉVIATIONS **1**

RÉSUMÉ **2**

CONTEXTE SCIENTIFIQUE **4**

I. INTELLIGENCE ARTIFICIELLE 4

1. Apprentissage automatique (AA) 4

a) Apprentissage non supervisé 5

b) Apprentissage par renforcement 5

c) Apprentissage supervisé *6*

2. Apprentissage profond 6

a) Le neurone artificiel 7

b) Les réseaux de neurones artificiels 8

3. Réseaux de neurones convolutifs 9

a) Principe de fonctionnement 9

b) Application des réseaux de neurones convolutifs profonds 11

c) Limites des réseaux de neurones convolutifs profonds 12

4. Apprentissage par transfert 12

5. Données 13

a) Les données d'entraînement 13

b) Les données de validation 13

c) Les données de test 14

6. Entraînement des données et notion d'hyperparamètre 14

a) Le taux d'apprentissage (learning rate) 14

b) Problème du sous-apprentissage et du sur-apprentissage 15

c) Autres hyperparamètres : patience, dimension de l'image 16

II. APPLICATION EN SANTÉ 18

1. De façon générale 18

2. Application en imagerie médicale 18

3. Application en imagerie musculo-squelettique 19

4. Spécificités de l'articulation du coude 20

a) Bases anatomiques 20

b) Fréquence des fractures occultes 21

5. Spécificités pédiatriques du coude 22

a) Bases en traumatologie pédiatrique *22*

b) Noyaux d'ossification du coude pédiatrique *22*

c) Repères normaux du coude pédiatrique *25*

III. JUSTIFICATIF DU TRAVAIL SCIENTIFIQUE 27

ARTICLE SCIENTIFIQUE **28**

I. ABSTRACT 29

II. INTRODUCTION 31
III. MATERIAL AND METHODS 32
1. Data collection 32
2. Reference standard and labelling 33
3. Training of the model 34
4. Image dimension 34
5. Data augmentation 35
6. Visualisation tool 35
7. Internal evaluation 35
8. External evaluation 36
9. Evaluation of radiologists 36
10. Statistical analysis 37
IV. RESULTS 38
1. Internal evaluation 38
2. External evaluation of model 39
3. Interaction with radiologists 40
4. Visualisation tool using GradCAM 44
V. DISCUSSION 46
VI. CONCLUSION 49
RÉFÉRENCES BIBLIOGRAPHIQUES 50

ABRÉVIATIONS

AA	Apprentissage Automatique
AUROC	Area Under the Receiver Operating Characteristic (Aire sous la courbe ROC)
DCNN	Deep Convolutional Neural Network (Réseaux de neurones convolutifs profonds)
DICOM	Digital Imaging and COmmunications in Medicine (Imagerie numérique et communications en médecine)
IA (AI)	Intelligence Artificielle (Artificial Intelligence)
IRM	Imagerie par Résonnance Magnétique
M1	Modèle 1
M2	Modèle 2
PACS	Picture Archiving and Communication System (Système d'archivage et de transmission d'images)
Se	Sensibilité
Sp	Spécificité
TDM	TomoDensitoMétrie

RÉSUMÉ

INTRODUCTION :

Les erreurs de diagnostic sur les radiographies traumatologiques sont courantes, notamment chez l'enfant. De nombreux modèles d'intelligence artificielle (IA) ont été développés pour détecter les lésions post-traumatiques mais peu ont été validés chez l'enfant. De plus, l'impact de ces modèles en pratique clinique est rarement évalué. L'objectif de cette étude était de développer un modèle d'IA capable de détecter les lésions post-traumatiques sur les radiographies du coude pédiatrique, d'évaluer ses performances puis l'impact de son utilisation par les radiologues.

MÉTHODE :

1956 radiographies pédiatriques du coude réalisées suite à un traumatisme ont été collectées chez 935 patients âgés de 0 à 18 ans, entre Janvier 2015 et Août 2019. Elles ont été réparties aléatoirement en trois ensembles de données : données d'entraînement, données de validation et de test interne. Les réseaux de neurones convolutifs profonds ont été formés en faisant varier les hyperparamètres (données d'entraînement et de validation) puis évalués sur les données de test interne. Les deux meilleurs modèles sélectionnés ont été évalués sur des données de test externe impliquant 120 patients, dont les radiographies ont été réalisées avec un équipement radiologique différent à une autre période (Juillet à Décembre 2014). Huit radiologues ont interprété ces données sans l'aide des modèles d'IA puis avec.

RÉSULTATS :

Deux modèles se sont démarqués sur les données de test interne. Le modèle 1 (M1) avait une précision de 95.8% et une AUROC de 0.983, avec le meilleur compromis sensibilité / spécificité (Se = 0.935 / Sp = 0.978). Le modèle 2 (M2) avait une précision de 90.5% et une AUROC de 0.975 avec une meilleure sensibilité mais une spécificité plus faible (Se = 0.974 / Sp = 0.844). Sur l'ensemble de données de test externe, M1 a conservé une bonne précision de 82.5% et AUROC de 0.916 (-0.067) ainsi qu'un bon compromis sensibilité / spécificité (Se = 0.847 / Sp = 0.803). En revanche, M2 a connu une baisse de performance avec une perte de précision jusqu'à 69.2% et d'AUROC à 0.793 (-0.182). M1 a significativement amélioré la sensibilité du radiologue (0.82 à 0.88 ; p = 0.016) et la précision (0.86 à 0.88 ; p = 0.047), tandis que M2 a significativement diminué la spécificité des lecteurs (0.86 à 0.83 ; p = 0.031).

CONCLUSION :

Le développement d'un modèle d'IA pour détecter les lésions post-traumatiques du coude pédiatrique en radiographie est faisable. Cependant, des modèles avec des performances proches *in silico* peuvent conduire les radiologues à améliorer (M1) ou à baisser (M2) leurs performances en pratique clinique, soulignant la nécessité d'une évaluation clinique précise des outils basés sur l'IA.

CONTEXTE SCIENTIFIQUE

I. INTELLIGENCE ARTIFICIELLE

L'intelligence artificielle (IA) correspond à l'ensemble des théories et techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine (1). Elle consiste donc à produire des machines capables de réaliser des actions que seul l'humain est censé pouvoir réaliser : apprendre, comprendre une langue, reconnaitre des images, résoudre un problème.

Alan Turing est l'un des premiers à publier sur le sujet lorsqu'il présente son « *Imitation Game* » en 1950, désormais appelé « test de Turing ». Ce test consiste à évaluer la capacité d'une machine à imiter la conversation humaine. Un être humain engage une conversation avec un autre humain et un ordinateur, en aveugle sans savoir à qui il s'adresse et le test est positif si l'humain n'est pas en mesure de discerner à qui il est en train de s'adresser (2).

1. Apprentissage automatique (AA)

En 1959, le Professeur SAMUEL définit l'apprentissage automatique (*machine learning*) comme la branche d'étude qui donne la capacité à un ordinateur d'apprendre à réaliser une tâche, sans l'avoir explicitement programmé au préalable (3).

L'AA consiste à créer des algorithmes capables d'apprendre et de faire des prédictions sans en programmer toutes les étapes. Ces algorithmes utilisent le mécanisme d'induction à partir des données puis le processus de généralisation pour obtenir une connaissance approchée de la connaissance exacte.

Il existe **trois systèmes d'apprentissage automatique** : non supervisé, par renforcement et supervisé (4).

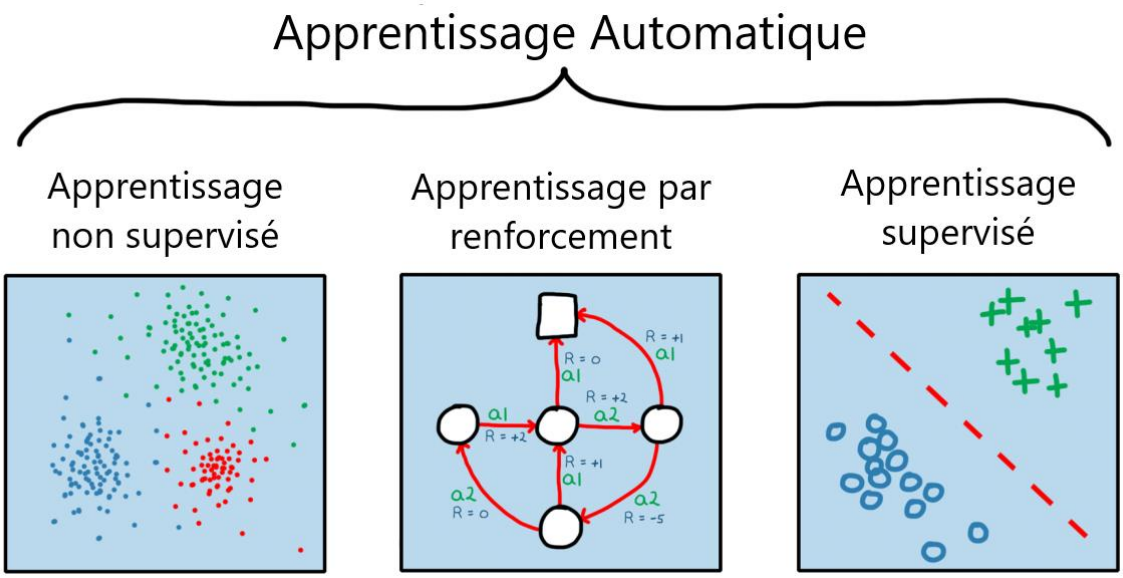

Figure 1 : Les différents types d'apprentissage automatique (adapté de (5)).

a) Apprentissage non supervisé

L'objectif de l'**apprentissage non supervisé** est de regrouper les données en fonction de leur ressemblance (6). Les données ne sont donc pas labellisées *a priori* et l'algorithme réalise une séparation des données en leur trouvant des points communs afin de les réunir en grappes (*clusters*) : c'est du **regroupement**.

Il s'agit donc de trouver des structures sous-jacentes à partir de données non étiquetées. Il est à ce jour surtout utilisé à visée exploratoire, afin de trouver des *patterns* communs entre plusieurs jeux de données.

b) Apprentissage par renforcement

L'**apprentissage par renforcement** consiste pour un modèle à apprendre à partir des interactions avec l'environnement. L'algorithme reçoit, à chaque interaction avec l'environnement, un signal rétro-actif (récompense ou pénalité), afin d'adapter sa décision lors d'une prochaine expérience similaire.

Il s'agit donc d'apprendre à partir d'expériences en optimisant une récompense obtenue après une séquence de décisions, comme le programme *Alpha-Go*. En 2015, il devient le premier programme à battre un joueur professionnel au jeu de Go après avoir réalisé des millions de parties d'entraînement (7).

c) Apprentissage supervisé

La majorité de l'AA employé couramment aujourd'hui utilise l'**apprentissage supervisé**, qui consiste à apprendre à partir de données labellisées. Le modèle apprend à classer un nouvel élément parmi un ensemble de classes pré-définies : c'est de la **classification**. Il est surtout utilisé pour essayer de prédire et classifier des données futures de façon la plus précise possible.

C'est ce type d'AA qui est de loin le plus utilisé en médecine et notamment une de ses composantes appelée apprentissage profond (*deep learning*).

2. Apprentissage profond

L'apprentissage profond (*deep learning*) est un sous-ensemble de l'apprentissage supervisé utilisant des réseaux de neurones profonds, c'est-à-dire des réseaux comprenant plusieurs couches de neurones, pour faire un modèle capable de simuler le fonctionnement neuronal humain.

a) Le neurone artificiel

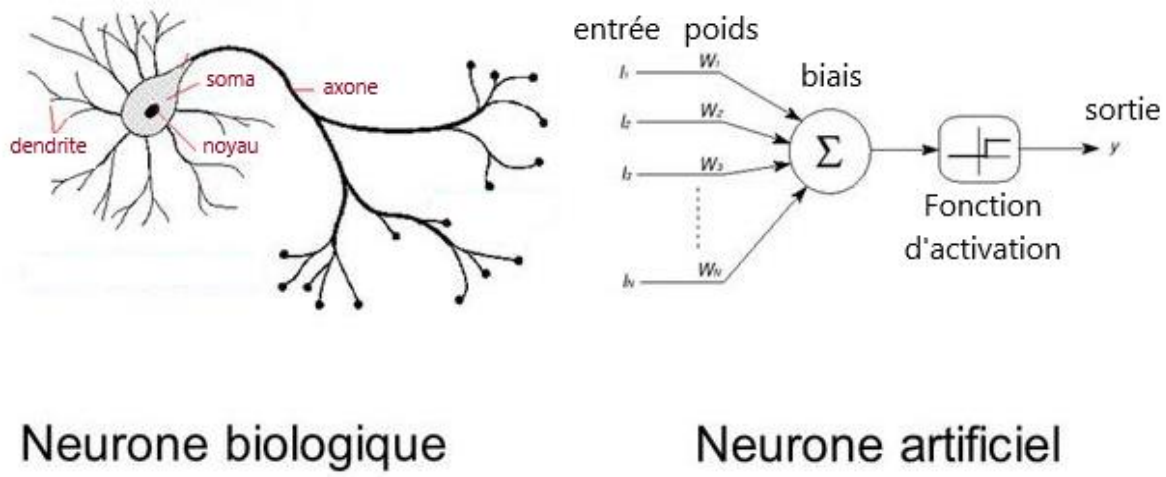

Figure 2 : Neurone biologique versus neurone artificiel.

A l'image du neurone biologique, les neurones artificiels utilisent des données en entrée et envoient un signal en sortie (adapté de (8)).

Dans le cerveau humain, le neurone reçoit des signaux chimiques et électriques provenant des synapses des multiples dendrites afférents qu'il intègre dans son corps cellulaire (*soma*). Lorsqu'un seuil d'excitation est dépassé, le neurone émet un signal d'activation avec les neurones de voisinage via son axone efférent. Comme le neurone biologique, le neurone artificiel est constitué de trois éléments :

- le **poids** : pondération des multiples entrées, à l'image des multiples dendrites du neurone biologique.

- le **biais** : valeur modificatrice des entrées, équivalent à l'intégration des signaux dans le corps cellulaire du neurone biologique.

- la **fonction d'activation** détermine la valeur numérique en sortie du neurone en fonction des valeurs en entrée, à l'image du signal d'activation émis par le neurone biologique.

Le poids et le biais ne sont pas connus par l'humain mais sont appris par le neurone artificiel lui-même à l'aide des données dans un processus d'apprentissage.

b) Les réseaux de neurones artificiels

Les réseaux de neurones artificiels sont constitués de multiples couches de neurones qui communiquent entre eux par des connexions antérogrades et rétrogrades (9).

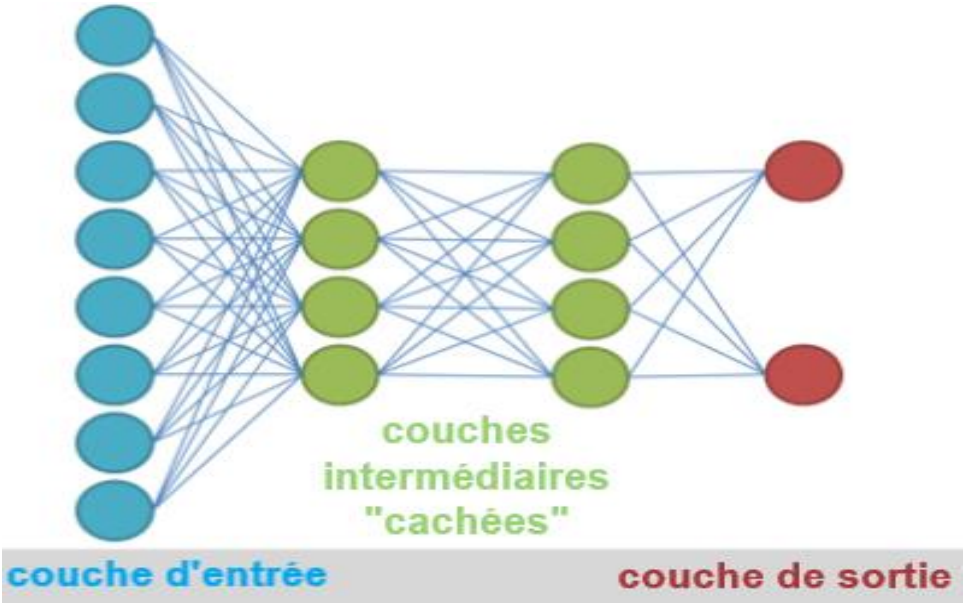

Figure 3 : Schématisation d'un réseau de neurones artificiels.

Chaque neurone est connecté à tous les neurones d'amont et d'aval. Ces réseaux sont constitués d'une couche d'entrée, de multiples couches intermédiaires et d'une couche de sortie. Les couches intermédiaires sont appelées couches « cachées » en raison du fait qu'elles ne produisent pas directement des sorties visibles mais calculent plutôt des représentations intermédiaires nécessaires au processus d'inférence (9). La couche d'entrée peut par exemple correspondre aux pixels d'une image et la couche de sortie à la prédiction de l'image.

Durant la phase d'apprentissage, le neurone évolue ensuite grâce à une série de rétro-propagation (*back-propagation*) : la prédiction du réseau est comparée aux valeurs théoriques, afin que les pondérations des neurones soient modifiées pour progressivement optimiser les résultats.

Les réseaux de neurones artificiels « classiques » ne sont pas optimisés pour étudier des images car elles constituent un nombre volumineux de données : dans ce cas, on privilégie l'utilisation des réseaux de neurones convolutifs.

3. Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs sont un type de réseaux de neurones artificiels spécialement conçu pour traiter des images. Contrairement aux réseaux de neurones artificiels habituels, les données ne sont pas préalablement labellisées par l'humain mais ce sont les réseaux de neurones eux-mêmes qui déterminent les caractéristiques de l'image.

a) Principe de fonctionnement

Ce type de réseau est classiquement composé de deux parties (de façon implicite) :

- Les premières couches calculent une représentation des données ; leur principe est donc d'analyser l'image en caractéristiques (*features*) ; c'est le rôle des couches de convolution qui le font sans l'intervention humaine, contrairement aux réseaux de neurones artificiels classiques.
- Les dernières couches déterminent la classe en fonction de cette représentation : c'est l'étape de sous-échantillonnage, effectuée par les couches de

regroupement. Elles consistent à réduire la taille de l'image et donc la quantité de paramètres à analyser tout en conservant les caractéristiques importantes de l'image.

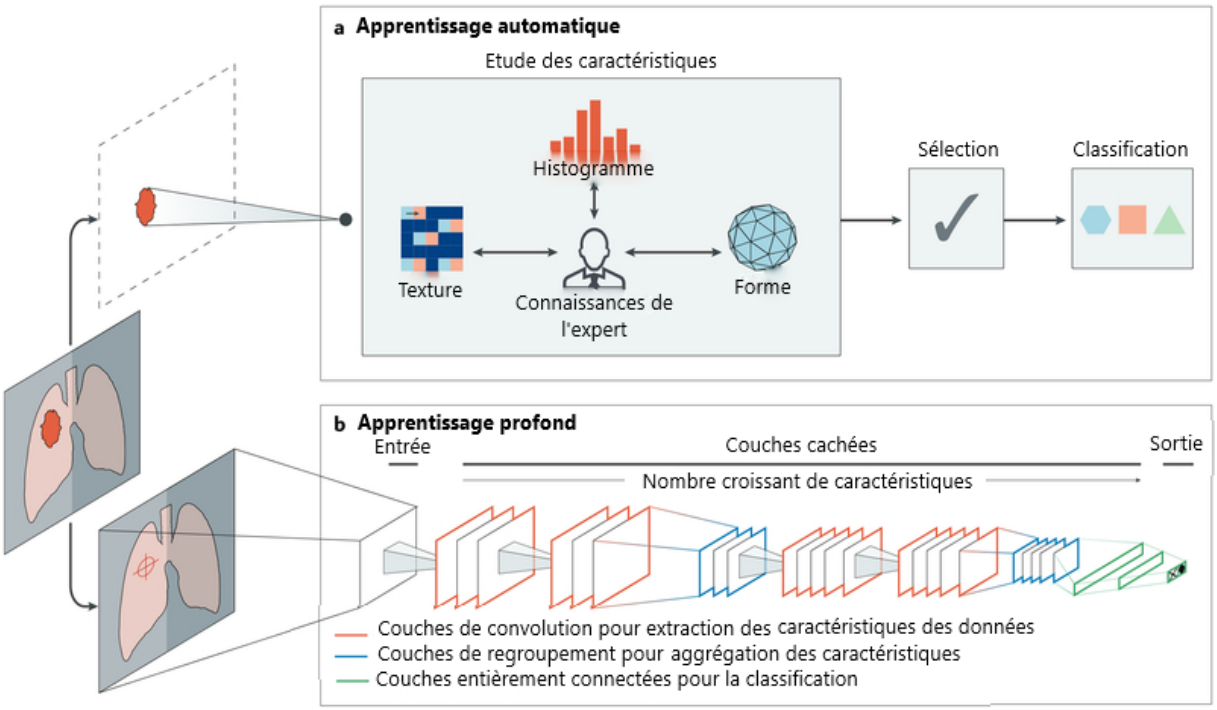

Figure 4 : Les deux principales méthodes d'apprentissage supervisé : Apprentissage automatique « standard » et apprentissage « profond » (adapté de (10)). L'apprentissage profond utilise des couches de convolution (*convolution layers*), qui déterminent les caractéristiques de l'image. Ensuite, les couches de regroupement (*pooling layers*) permettent de proposer une classification.

Ces réseaux sont largement utilisés en radiologie car ils sont capables de considérer de nombreuses caractéristiques propres à l'image sans qu'elles ne soient formalisées *a priori*.

Ces progrès technologiques puissants ne sont pas récents car leurs fondements datent d'il y a 30 ans mais leurs applications sont restées en dormance jusqu'en 2012.

b) Application des réseaux de neurones convolutifs profonds

ImageNet est une base de données de plusieurs millions d'images annotées à destination des travaux de recherche (11). Depuis 2010, le projet ImageNet organise un concours annuel de reconnaissance d'images dont le but est de développer un modèle capable de détecter et classifier précisément des objets dans les images naturelles (12).

Figure 5 : Données tirées d'ImageNet regroupant des millions d'images (13).

Jusqu'en 2012, les participants aux concours utilisaient l'apprentissage automatique (*machine learning)* classique avec un taux d'erreur qui stagnait autour de 25% (non compatible avec l'imagerie médicale).

En 2012, Y. Lecun, G. Hinton et Y. Bengio ont largement dominé leurs concurrents en utilisant l'apprentissage profond (*deep learning)* avec réseaux de neurones convolutifs et ont ainsi fait chuter le taux d'erreur à 16% (14).

Depuis, les performances s'améliorent chaque année avec maintenant un taux d'erreur de moins de 5%.

c) Limites des réseaux de neurones convolutifs profonds

Cependant, le modèle propose un résultat mais il n'est pas possible de comprendre formellement à partir de quel(s) argument(s) a été réalisée la prédiction : on ne peut donc pas contrôler objectivement le résultat. Il s'agit donc d'une « boîte noire » car il faut accepter les résultats sans véritable justification scientifique, ce qui va à l'encontre de la démarche scientifique historique.

On peut tenter d'approcher une compréhension de cette boîte noire grâce à des cartes de chaleur (*heat-map*) (15). Ces cartographies consistent à mettre en évidence les pondérations qui ont le plus participé au résultat de l'algorithme, en faisant apparaitre ces zones de l'image selon un gradient de couleur (par exemple en rouge, versus bleu pour les zones ne participant pas ou peu à la performance de l'algorithme) (16).

4. Apprentissage par transfert

Il s'agit d'une approche d'apprentissage automatique qui adapte et transfère les connaissances acquises d'un modèle lors d'une tâche précédente à une nouvelle tâche différente mais apparentée (17).

Par exemple, un réseau de neurones convolutifs pré-entraîné sur ImageNet pour la reconnaissance visuelle d'images non médicales a été utilisé pour la segmentation des gliomes cérébraux en imagerie par résonance magnétique (IRM) (18) ou encore pour la prédiction de la survie des patients atteints d'adénocarcinome pulmonaire en imagerie tomodensitométrique (TDM) (19).

5. Données

Les données sont classiquement divisées en **trois sous-ensembles** : les données d'entraînement, les données de validation et les données de test.

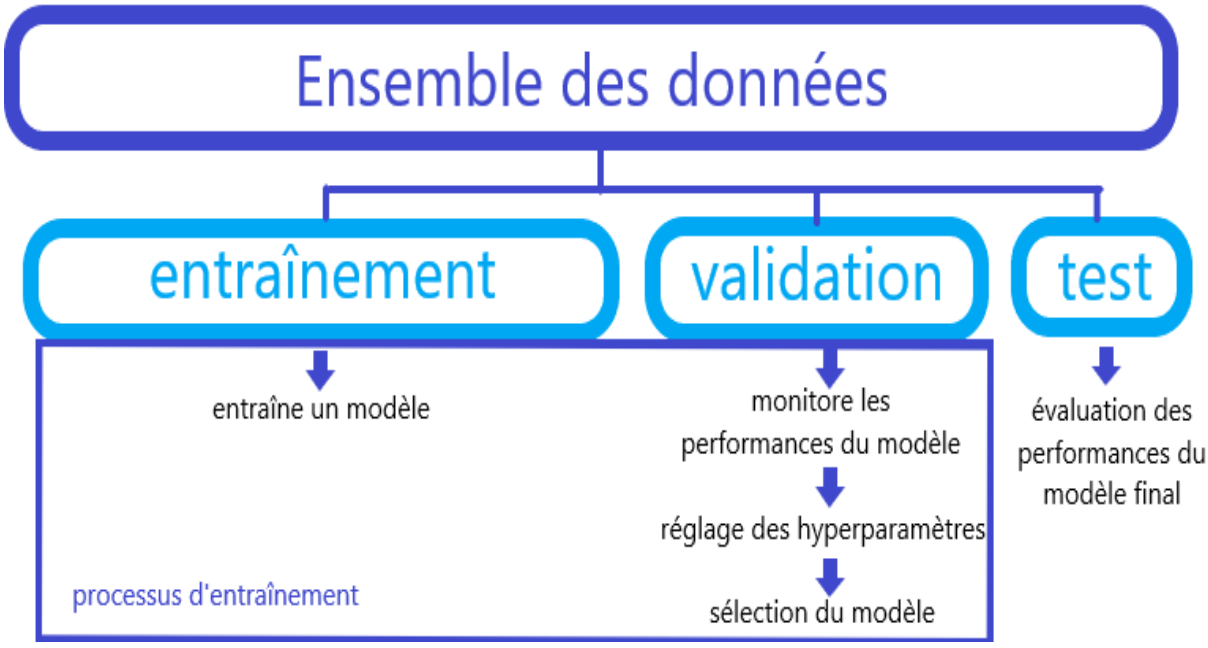

Figure 6 : Représentation des données, divisées en trois sous-ensembles.

a) Les données d'entraînement

Elles permettent d'entraîner le modèle et donc de réaliser l'apprentissage de l'algorithme. Elles doivent être représentatives de la cible et les classes doivent idéalement être équilibrées.

b) Les données de validation

Elles évaluent de façon primaire le modèle à l'issue de la phase d'entraînement. Elles sont utilisées pour monitorer les performances du modèle et les affiner en variant les hyperparamètres afin de sélectionner le modèle le plus fiable. A l'issue de cette phase, le modèle est gelé pour évaluation.

c) Les données de test

Elles évaluent les performances du modèle final. Les bonnes pratiques veulent que les données incluses dans le set de test ne soient pas connues par l'algorithme (présentes ni dans le set d'entraînement ni dans le set de validation) voire proviennent de données extérieures. Une fois le set de test utilisé, il n'y a plus de modification possible de l'algorithme.

6. Entraînement des données et notion d'hyperparamètre

La création d'un réseau de neurones convolutifs nécessite la création de centaines de modèles faisant varier de nombreux **hyperparamètres** : le taux d'apprentissage, les dimensions de l'image, etc.

Un hyperparamètre est un paramètre dont la valeur est utilisée pour contrôler le processus d'apprentissage. Choisir les meilleures valeurs pour les hyperparamètres permettent d'améliorer significativement les performances d'un modèle.

a) Le taux d'apprentissage (learning rate)

Le taux d'apprentissage correspond à la taille du pas à chaque itération ; c'est un facteur crucial en *machine learning* (20). Quand on corrige petit à petit les poids d'un réseau de neurones, le taux d'apprentissage contraint l'amplitude de la correction. Si le pas est grand, les corrections vont être grandes et il est possible qu'on saute au-dessus de la valeur optimale du paramètre alors l'algorithme diverge. Si le pas est très petit, l'algorithme va avancer très doucement vers la valeur optimale et peut quand même diverger lors de l'approche de la valeur optimale.

Pour éviter cela, il y a deux solutions. Soit l'algorithme est brutalement arrêté au bout d'un certain temps lorsqu'on est satisfait du paramétrage obtenu : c'est l'arrêt

précoce (*early stopping*) (21). Soit le taux d'apprentissage est diminué petit à petit pour faire des pas de plus en plus petits pour ne pas diverger.

b) Problème du sous-apprentissage et du sur-apprentissage

Le sous-apprentissage et le sur-apprentissage sont liés à la complexité du modèle utilisé par rapport à la complexité de la structure à exhiber dans les données. En cas de sous-apprentissage (*underfitting*), le modèle sera trop simple donc non exploitable. En cas de sur-apprentissage (*overfitting*), le modèle ne sera pas généralisable et donc non exploitable. Il s'agit d'un perpétuel compromis entre précision et généralisabilité.

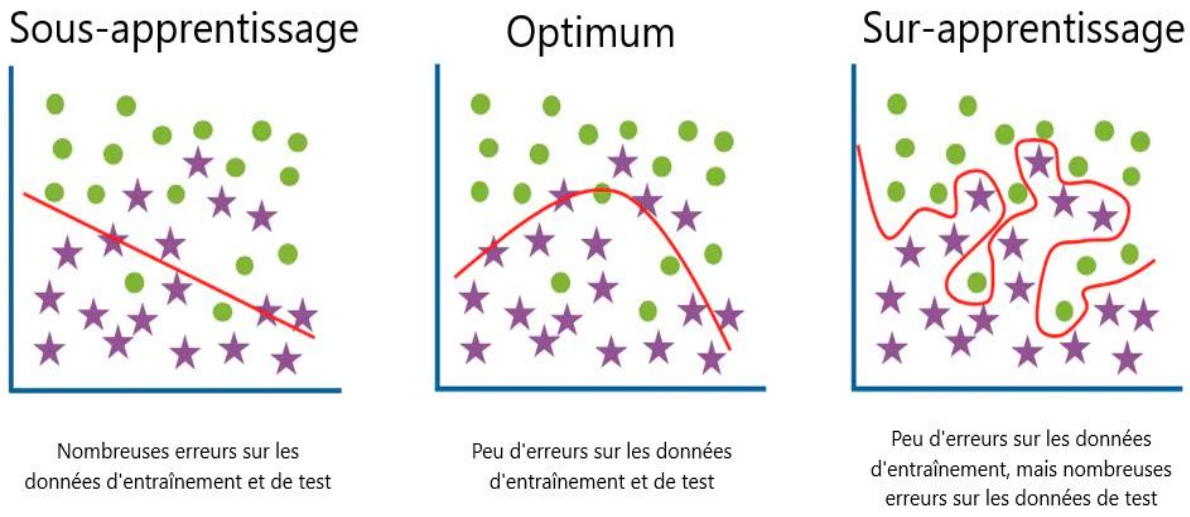

Figure 7 : Schématisation du sous-apprentissage et du sur-apprentissage (22).

En cas de sous-apprentissage (*underfitting*) : le modèle oublie de nombreuses étoiles, il fait beaucoup d'erreurs sur les données d'entraînement et donc de test. Concernant le sur-apprentissage (*overfitting*), le modèle reconnaît toutes les étoiles : il ne fait pas d'erreur sur les données d'entraînement qu'il connaît parfaitement mais fera de nombreuses erreurs sur les données de test qui seront différentes : il n'est pas généralisable.

c) Autres hyperparamètres : patience, dimension de l'image

La **patience** correspond au nombre d'époques. Une époque correspond à un unique apprentissage sur toutes les données. L'ensemble du jeu de données est donc itéré à chaque époque. Plus le nombre d'époques à réaliser est grand, meilleure devrait être la précision, mais cela peut conduire également au sur-apprentissage. C'est pour cette raison que si le modèle ne s'améliore pas sur un certain nombre d'époques, l'entraînement s'arrête.

La **dimension de l'image** correspond au produit de la largeur et de la hauteur de l'image. Par exemple, une image de 512x512 (largeur x hauteur) contient un total de 262 144 pixels. Un pixel correspond à chaque point d'une image électronique. Les performances du modèle varient en fonction de la dimension de l'image (23).

Le nombre d'images utilisées pour entrainer le réseau qui correspond à la **taille du lot** (*batch size*) est essentielle. Afin d'augmenter le nombre d'images labellisées, le principe d'augmentation des données (*data augmentation*) consiste à augmenter artificiellement le nombre d'images en y appliquant des transformations, telles qu'une rotation de l'image, du flou ou du bruit introduit dans l'image ou encore le changement de luminosité de l'image. On augmente ainsi la diversité et le champ d'apprentissage du modèle (24).

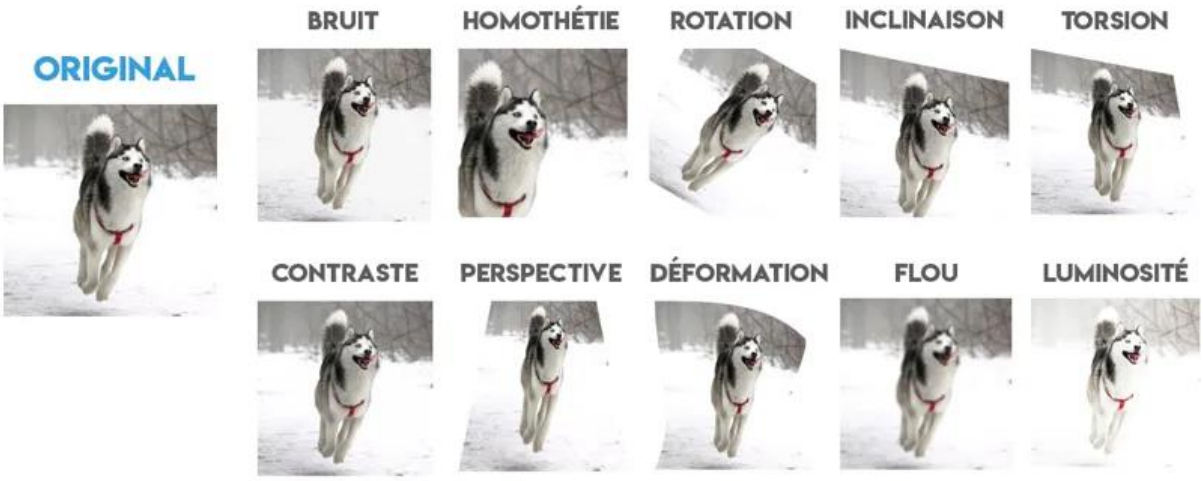

Figure 8 : Principales techniques d'augmentation des données (d'après (24)).

II. APPLICATION EN SANTÉ

1. De façon générale

Le nombre d'articles portant sur l'IA dans le domaine de la santé a augmenté de façon exponentielle (25). Les avancées spectaculaires récentes dans ce domaine sont surtout liées à la démocratisation des ordinateurs et cartes graphiques, plus accessibles de par leur coût et qui ont des puissances de calcul de plus en plus importantes.

En parallèle, le développement récent du « *Big data* » permet de constituer une source exponentielle de données disponibles et facilement accessibles pour entraîner ces modèles. Il faut corréler à cela les investissements financiers majeurs récents dans ces technologies, notamment dans le secteur de la santé (26).

De plus en plus de radiologues s'intéressent donc à l'IA et surtout à l'aide qu'elle pourrait apporter au quotidien (27).

2. Application en imagerie médicale

L'IA en radiologie pourrait constituer une aide au diagnostic dans la détection des lésions, leur classification et la segmentation anatomique (28). Cependant l'aide au diagnostic représente une infime partie de l'aide apportée par l'IA (29).

En effet, l'IA pourrait optimiser le temps médical (30), faciliter la gestion du flux des patients (31), ainsi que l'acquisition des images notamment avec réduction des doses (32). L'IA permettrait également d'optimiser l'interprétation en alignant les différents plans de coupe ou en synchronisant les examens antérieurs (33).

L'IA pourrait également prédire le caractère suspect d'une lésion ou la réponse au traitement d'une lésion en fonction des caractéristiques de l'image à l'aide de la radiomique (34,35).

3. Application en imagerie musculo-squelettique

La détection des fractures constitue un enjeu en imagerie musculo-squelettique. En effet, les fractures non diagnostiquées représentent plus de 80% des erreurs de diagnostic aux urgences (36).

La première publication dans ce domaine remonte à 2017 : cette étude montrait l'efficacité d'algorithmes usuels de *machine learning* dans la détection de paramètres de base des radiographies : latéralité, région anatomique, type de cliché avec une efficacité de 90% et une précision de 83% concernant la détection des fractures en prenant pour référence des chirurgiens orthopédistes (37).

D'autres études ont ensuite démontré la faisabilité de la détection de fractures du poignet (38), de l'extrémité proximale de l'humérus (39), fractures pertrochantériennes de la hanche (40), de la cheville (41) ou du rachis (42). Dans ce type d'études, les faux négatifs comprennent les fractures sans déplacement significatif. Les faux positifs concernent les clichés réalisés avec plâtre et les enfants (liés au cartilage de croissance) (43).

L'utilisation de cartes de classes d'activation, avec notamment des cartes de chaleur permet d'essayer de comprendre où se situe la zone suspectée pathologique et d'orienter le radiologue dans l'interprétation des images : elles peuvent être très utiles dans la détection des fractures.

4. Spécificités de l‘articulation du coude

a) Bases anatomiques

Le coude est constitué de trois articulations : huméro-radiale (par le biais du capitulum), huméro-ulnaire (par le biais de la trochlée) et radio-ulnaire proximale.

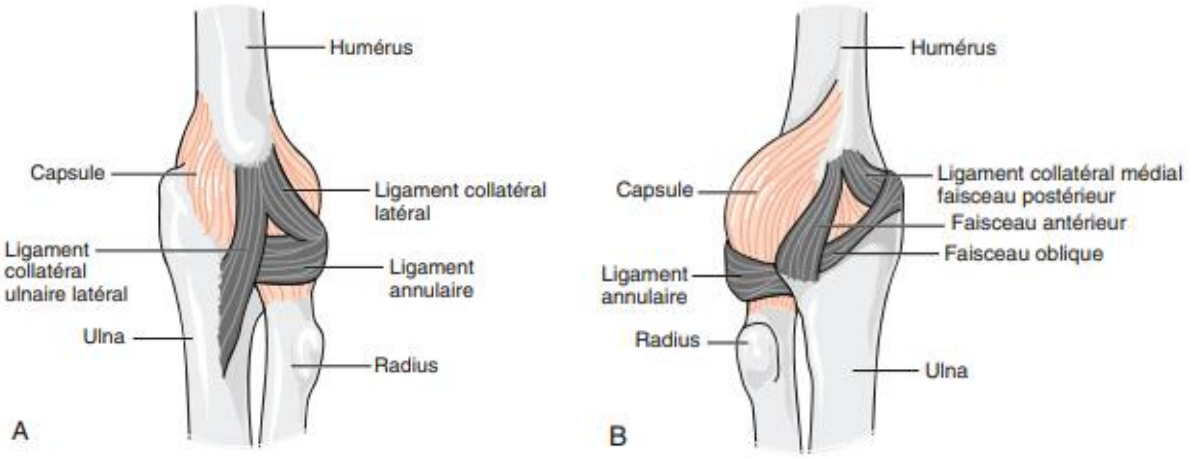

Figure 8 : Schématisation de vue latérale (A) et de vue médiale (B) du coude (d’après (44)). Le complexe ligamentaire latéral comprend le ligament collatéral radial, le ligament ulnaire latéral et le ligament annulaire. Le complexe ligamentaire médial appelé ligament collatéral ulnaire est composé de trois faisceaux.

La stabilité du coude est assurée par un trépied principal (l'articulation huméro-ulnaire avec le processus coronoïde et les ligaments collatéraux médial et latéral). Si l'un de ces éléments est lésé, il peut être compensé par l'un des éléments du trépied secondaire (tête radiale, capsule articulaire et muscles) (45). Les muscles fléchisseurs (brachial antérieur et biceps brachial) et les extenseurs du coude (triceps brachial) maintiennent une bonne coaptation des pièces osseuses.

Le bilan radiographique en cas de traumatisme repose sur les clichés de face et de profil.

b) Fréquence des fractures occultes

Les fractures occultes du coude sont classiques, et peuvent ne se manifester que par un épanchement articulaire isolé en radiographie. Il faut donc rechercher systématiquement l'épanchement articulaire qui doit motiver la réalisation d'examens complémentaires à la recherche d'une fracture radiographiquement non décelable (échographie, scanner, IRM). (46).

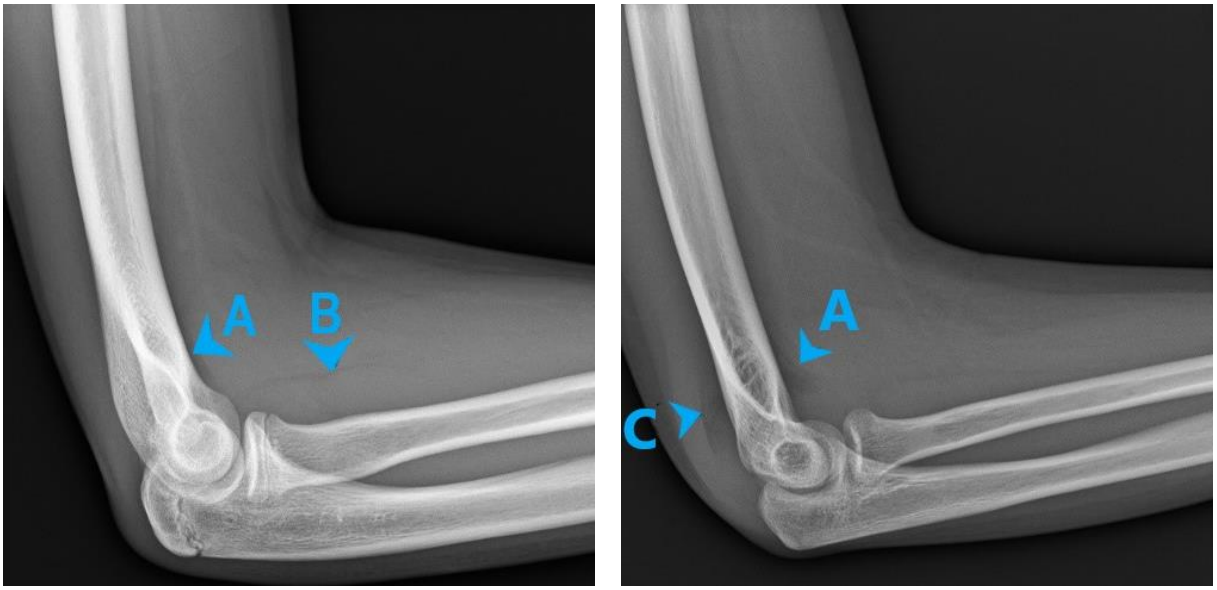

Figure 9 : Radiographies normale (à gauche) et pathologique (à droite).

A : liseré coronoïdien / antérieur.

B : liseré court supinateur : situé en avant de l'extrémité supérieure du radius, entre le court supinateur en arrière, le rond pronateur et long supinateur en avant : c'est un bon indicateur pour les traumatismes de la tête radiale.

C : liseré de la fossette olécranienne / postérieur (visible uniquement en cas d'épanchement abondant).

En cas d'épanchement articulaire, le liseré antérieur (coronoïdien) est refoulé et amputé, prenant une forme triangulaire. Si l'épanchement est important, on visualise le liseré postérieur (fossette olécranienne) (47).

5. Spécificités pédiatriques du coude

a) Bases en traumatologie pédiatrique

Certaines fractures sont spécifiques à l'enfant en raison de la plasticité des structures osseuses, notamment les fractures en motte de beurre, fractures en bois vert, ou encore fractures-déformations plastiques.

Il existe également des fractures du cartilage de croissance : dans ce cas, la classification de Salter et Harris permet d'évaluer le décollement épiphysaire. Elle est essentielle pour évaluer la croissance ultérieure de l'os, en mesurant notamment le risque d'épiphysiodèse.

La majorité des fractures de la palette humérale chez l'enfant sont les fractures supracondyliennes, qui représentent environ 50% des fractures du coude de l'enfant (48). Le principal signe radiographique de ces fractures est l'hémarthrose, qui doit faire suspecter le diagnostic de fracture articulaire, même si le trait fracturaire n'est pas visible. Chez l'enfant, les fractures de la tête radiale, de l'olécrâne et du processus coronoide sont rares.

La pronation douloureuse est la plus fréquente des lésions ostéo-articulaires du coude chez le petit enfant (49). Elle survient dans 98 % des cas avant 6 ans et est liée à la luxation du ligament annulaire au sein de l'interligne huméro-radial. Dans ce cas, le bilan radiographique n'a aucun intérêt car il est tout à fait normal, notamment il n'y a pas d'hémarthrose (50).

b) Noyaux d'ossification du coude pédiatrique

Le coude de l'enfant présente six noyaux d'ossification secondaire qui apparaissent et fusionnent dans un ordre reproductible d'un enfant à l'autre (ossification de type enchondrale).

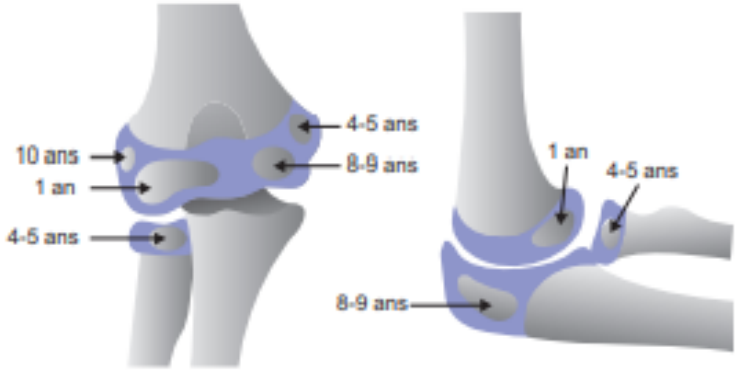

Figure 10 : Ordre d'apparition des noyaux d'ossification avec pour moyen mnémotechnique « CRITOL » : Capitulum (à 1 an), puis la tête Radiale et l'épicondyle Interne (vers 4-5 ans), ensuite la Trochlée et l'Olécrane (vers 8-9 ans) et pour finir l'épicondyle Latéral (vers 10 ans) d'après (50).

Cet élément est essentiel à connaître pour ne pas confondre une fracture et un noyau d'ossification. Les noyaux d'ossification du capitulum, de la trochlée et de l'épicondyle latéral fusionnent entre eux vers 10 - 12 ans, pour ensuite fusionner avec la métaphyse vers 13 ans. Le noyau de l'épicondyle médial est le dernier à fusionner vers 14 ans chez la fille et 17 ans chez le garçon. Certains noyaux d'ossification peuvent être bipartite, notamment le noyau olécranien et ne doivent alors pas être confondus avec une fracture (50).

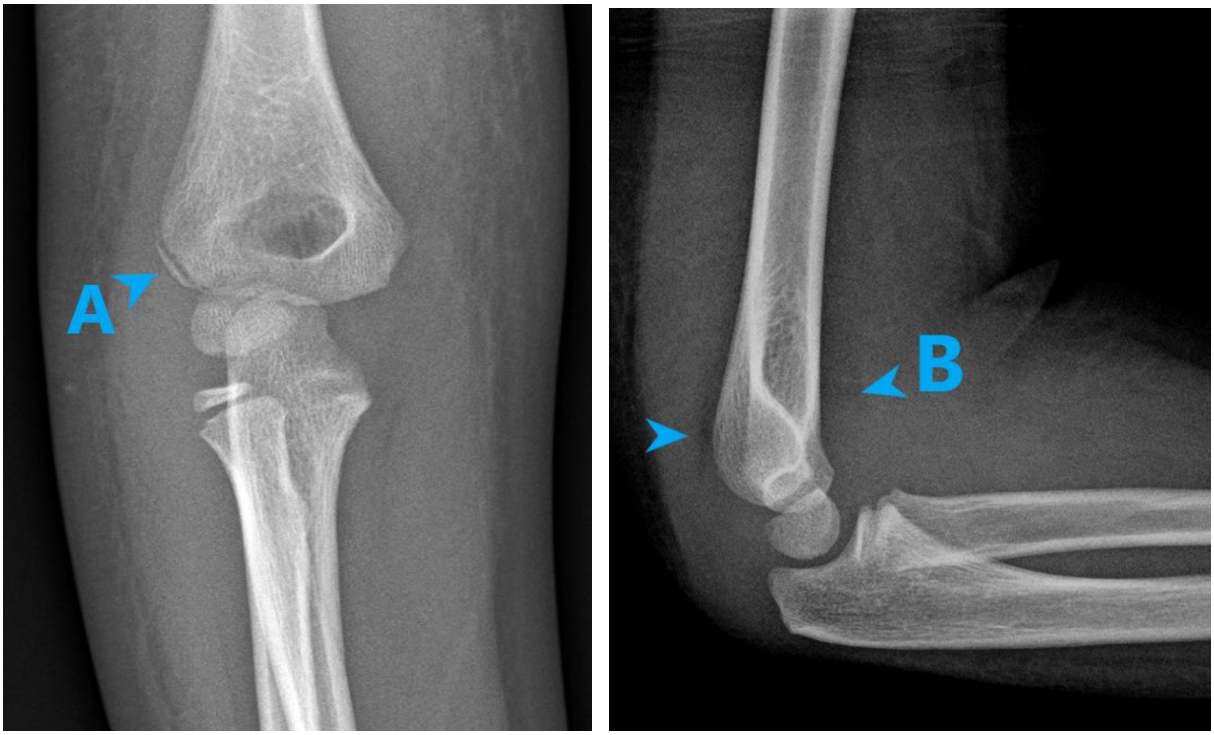

Figure 11 : Radiographies de face (à gauche) et de profil (à droite).

On visualise une fracture du condyle latéral (A) chez un enfant de 4 ans pouvant être confondue avec le noyau épicondylien latéral non fusionné (or il n'apparaît que vers 10 ans). On note par ailleurs l'épanchement articulaire sur le cliché de profil (B).

L'analyse des tissus mous est également indispensable car la fracture-avulsion d'un noyau cartilagineux (non encore ossifié) ne peut se manifester que par une infiltration des tissus mous en radiographie.

Figure 12 : Radiographie du coude de face chez un enfant de 1 an.

On retrouve une nette infiltration des tissus mous de l'épicondyle médial évocateur d'une fracture-avulsion du noyau cartilagineux épicondylien médial (pas encore ossifié), confirmé en échographie.

c) Repères normaux du coude pédiatrique

Chez l'enfant, l'analyse des lignes est essentielle pour s'assurer de l'absence de luxation ou malposition d'un noyau d'ossification.

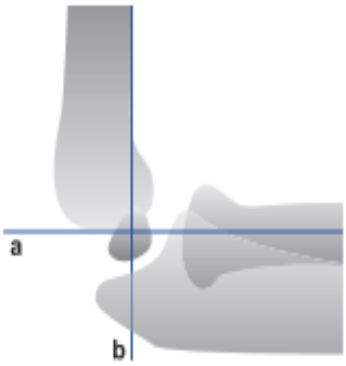

Figure 13 : Lignes essentielles à étudier lors de l'analyse d'une radiographie de profil d'un coude de l'enfant (d'après (50)).

La ligne capitulo-radiale (a) qui correspond à l'axe de la diaphyse radiale qui passe par le centre du capitulum de face et de profil, quel que soit le degré de flexion du coude (50). La ligne humérale antérieure (b) qui correspond à la tangente à la corticale antérieure de l'humérus. Elle passe par le tiers moyen du noyau du capitulum.

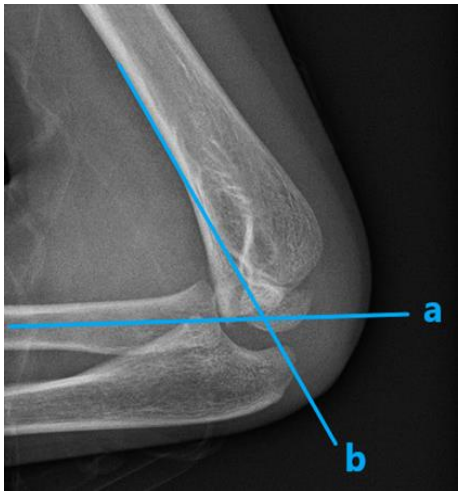

Figure 14 : Radiographie du coude de profil chez un enfant de 7 ans qui présente un défaut d'extension du coude.

La ligne capitulo-radiale (a) ne passe plus par le centre du capitulum et la ligne humérale antérieure (b) ne passe plus par le tiers moyen du capitulum mais par le tiers antérieur.

III. JUSTIFICATIF DU TRAVAIL SCIENTIFIQUE

Les traumatismes du membre supérieur sont un motif fréquent de consultation aux urgences et nécessitent généralement un bilan radiographique (51).

L'analyse d'une radiographie du coude chez l'enfant nécessite l'étude osseuse à la recherche de fracture, l'étude des liserés graisseux à la recherche d'épanchement articulaire et également l'étude des parties molles de façon générale à la recherche d'une fracture-avulsion d'un noyau d'ossification.

Le taux d'erreur de diagnostic le plus important en ce qui concerne les fractures du membre supérieur chez l'enfant se situe au coude (77 % des erreurs de diagnostic) (52,53). Les erreurs de diagnostic sont plus fréquentes chez l'enfant en raison des noyaux d'ossification et de la variation de l'ossification (48,54).

La majorité des modèles d'intelligence artificielle qui permettent de détecter les fractures en radiographie ne sont pas validés chez l'enfant (55). De plus, malgré les nombreux articles de la littérature, l'intérêt est focalisé sur les performances *in silico* des algorithmes, mais rarement sur les performances issues de l'interaction entre l'humain et l'algorithme utilisé (56). Enfin, des travaux récents sur cette thématique commencent à montrer l'impact potentiellement délétère d'un modèle pourtant performant *in silico* sur les diagnostics portés par les médecins (57).

L'objectif principal de cette étude a été de créer un algorithme d'intelligence artificielle capable de détecter une lésion traumatique du coude chez l'enfant en radiographie. Secondairement, cet algorithme a été soumis à des radiologues sur un échantillon de radiographies de coude pédiatrique pour connaître l'impact de l'utilisation de ces modèles en pratique clinique.

ARTICLE SCIENTIFIQUE

FROM DATA TO DIAGNOSIS: DEVELOPMENT AND EVALUATION OF AN ARTIFICIAL INTELLIGENCE MODEL FOR THE DETECTION OF POST-TRAUMATIC INJURIES ON ELBOW RADIOGRAPHS IN CHILDREN.

Clémence ROZWAG [1,2], Franck VALENTINI [3], Anne COTTEN [1,2], Xavier DEMONDION [1,2], Philippe PREUX [1,3,4], Thibaut JACQUES [1,2]

1 : Lille University, Lille, France

2 : Lille University Hospital Center, Lille, France

3 : Inria Lille – Nord Europe, équipe SequeL

4 : UMR CNRS 9189

I. ABSTRACT

OBJECTIVE:

Diagnostic errors are common on trauma x-rays, especially in children. Many artificial intelligence models have been developed to detect post-traumatic lesions, but few have been validated in children. In addition, the impact of these models on radiologists' interpretation has rarely been assessed. The objective of this study was to develop a model able to detect post-traumatic injuries on pediatric elbow x-rays using artificial intelligence (AI), then to evaluate its performances and to assess its impact on radiologists' interpretation.

METHODS:

A total of 1956 pediatric elbow radiographs performed in the event of a trauma were retrospectively collected from a total of 935 patients aged from 0 to 18 years, between January, 2015 and August, 2019. These x-rays were randomly divided into three datasets: training set (72%), validation set (11%) and internal test set (17%). Deep convolutional neural networks (DCNN) were trained by varying hyperparameters (training and validation sets) then evaluated on the internal test set. Secondarily, the two best models were selected then evaluated on an external test set involving 120 patients, whose x-rays were performed on a different radiological equipment in another time period (July to December, 2014). Eight radiologists were asked to interpret this external test set without the help of AI models then with.

RESULTS:

Two models stood out on the internal test set. Model 1 (M1) had an accuracy of 95.8% and an AUROC of 0.983 on the internal test set, with the best sensitivity/specificity compromise (Se=0.935 / Sp=0.978). Model 2 (M2) had an accuracy of 90.5% and an AUROC of 0.975 on the internal test set, with a better sensitivity but a lower specificity (Se=0.974 / Sp=0.844).

On the external test set, M1 kept a good accuracy of 82.5% and AUROC of 0.916 (-0,067) and also a good sensitivity/specificity compromise (Se=0.847 / Sp=0.803). On the other hand, M2 had a drop in performance with a loss of accuracy down to 69.2% and of AUROC to 0.793 (-0.182).

M1 significantly improved radiologist's sensitivity (0.82 to 0.88, p=0.016) and accuracy (0.86 to 0.88, p=0,047), while M2 significantly decreased specificity of readers (0.86 to 0.83, p=0.031).

CONCLUSION:

End-to-end development of a DCNN model to assess post-traumatic injuries on elbow x-ray in children was feasible and showed that models with close metrics *in silico* can lead radiologists to either improve (M1) or lower (M2) their performances in clinical settings, underlining the need for precise clinical evaluation of AI-based tools.

II. INTRODUCTION

Detection of fractures is an issue in musculoskeletal imaging. Indeed, missed fractures represent more than 80% of diagnostic errors in the emergency department (36).

Trauma of the upper limb frequently requires radiographic workup (51). The highest rate of diagnostic error in upper limb fractures in children lays at the elbow (77% of diagnostic errors) (52,53). Misdiagnosis is more common in children due to ossification centers and variation in ossification (48,54).

Most of the artificial intelligence models that can detect fractures on X-ray are not validated in children (55). Moreover, despite numerous articles in the literature, focus is generally made on the *in silico* performance of algorithms, but only rarely on the performance resulting from the interaction between humans and the algorithm (56). In addition, recent work on this topic tends to show the potentially deleterious impact of a model, though effective *in silico*, on the diagnoses made by doctors (57).

The purpose of this study was therefore to develop a model using artificial intelligence (AI) to detect post-traumatic injuries on pediatric elbow x-rays, and then to assess the impact of its use on radiologists' performances.

III. MATERIAL AND METHODS

1. Data collection

All elbow X-rays performed in patients aged 0 to 18 years between January 1, 2015 and August 31, 2019 in the emergency department of the University Hospital of Lille (France) were retrospectively gathered from the PACS (Intellispace PACS, Philips). Series with only one view or with imperfect profile x-ray view were not excluded, because of their frequency in children, especially after an upper limb trauma. However, examinations including radio-opaque elements such as surgical screws or casts were not included. A total of 1956 x-rays from 935 patients (485 male and 450 female) were collected and de-identified. The de-identification of data and their retrospective analysis was approved by the institutional board under the reference DEC19-279.

All these examinations were performed on the same x-ray device (Fujifilm, Tokyo, JP).

Radiographs were randomly divided into a training set (n=1411, 72%, 668 patients), a validation set (n=209, 11%, 99 patients) and an internal test set (n=336, 17%, 168 patients) (Figure 1). The age repartition of patients is summarized in Table 1.

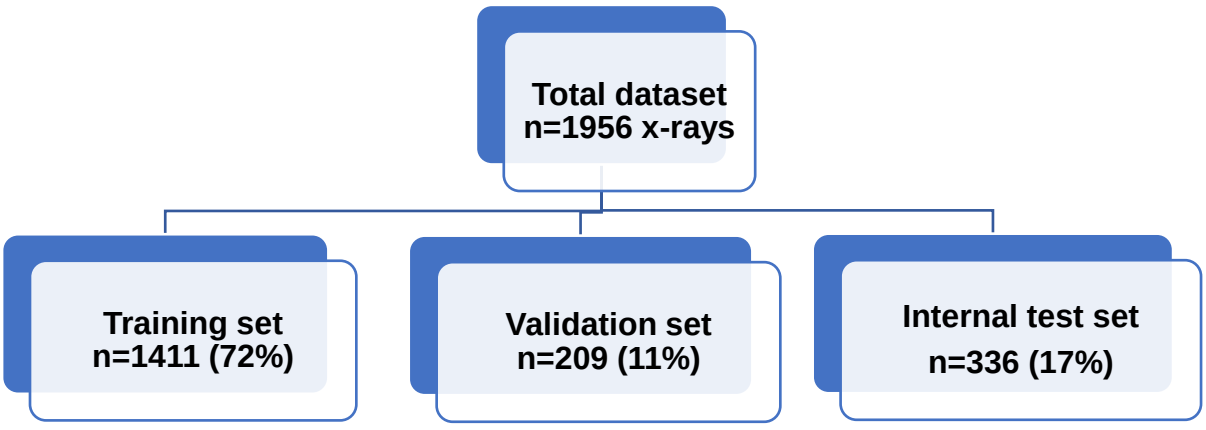

Figure 1: Repartition of the datasets.

Age (years)	[0-3[	[3-6[	[6-9[	[9-12[	[12-15[	[15-18]
n	79 (8.4%)	168 (18.0%)	181 (19.4%)	164 (17.5%)	146 (15.6%)	197 (21%)

Table 1: Age repartition in the 935 patients of the dataset. n: number of patients.

2. Reference standard and labelling

Two radiologists (one resident with 2 years of experience in trauma radiology, and a senior radiologist with 7 years of experience) classified in consensus all elbow X-rays in two groups: normal or pathological, depending on the findings. Examinations were labelled as pathological in the following situations: visible fracture, articular dislocation, soft tissues change of potential post-traumatic origin (*e.g.* fat pad sign or intra-articular fluid, even if no fracture was directly visible). A total of 1171 radiographs were considered normal and 785 pathological (Table 2).

	Pathological findings	**Normal examinations**
Training set	n=537 (38%)	n=874 (62%)
Validation set	n=92 (44%)	n=117 (56%)
Test set	n=156 (47%)	n=180 (53%)

Table 2: Repartition of examinations in the different datasets depending on the label. n: number of radiographs.

3. Training of the model

The training and validation set were used to optimize DCNN by modulating architecture, depth and hyperparameters. Models have been trained by varying three main hyperparameters: patience, learning rate and image dimensions. Table 3 shows different values of hyperparameters.

Hyperparameters	Values
Learning Rate	[0.00005, 0.00003, 0.00001, 0.000005, 0.000003, 0.000001]
Patience	[5, 10, 15, 20, 25] (learning rate was reduced by 50% if loss has not improved over N epochs)
Dimension	[224x224 / 512x512 / 1024x1024 / Crop224x224 / Crop512x512]

Table 3: different values of hyperparameters.

4. Image dimension

Originally, DICOM files had a dimension of 4096x4096 pixels. To reduce the computational cost of training the DCNN, this resolution was downscaled to 1024x1024 for each radiograph. On each image of dimension 1024x1024, radiologists identified an area containing the whole elbow joint (minimal size of 224x224), using a generic labelling tool (RectLabel, Mac Os) in order to obtain cropped images (Crop224x224). To obtain Crop512x512 X-rays, the original X-ray was downscaled to 2048x2048 and the bounding box from Crop224x224 was converted on this image in order to obtain the same area but with higher precision. An “early stopping” was added, to stop the training as soon as the model did not improve over 50 epochs.

5. Data augmentation

In order to increase the number of data, several rotations of the radiographs were performed. Vertical and horizontal transformations were randomly applied, and image rotation between -45° and + 45 ° was randomly used. The image normalization was performed using conventional ImageNet normalizations ([0.485, 0.456, 0.406] , [0.229, 0.224, 0.225]) since the DCNN was pre-trained using ImageNet.

6. Visualisation tool

GradCAM algorithm was implemented to visualize the class activation heatmap of the algorithm. The scale of the GradCam varied from blue (low) to red (high) depending on the weight of each area in the final output of the algorithm.

7. Internal evaluation

Internal performances of the model were measured on the internal test set which was composed of 336 x-rays, included neither in training nor in validation sets, from 168 patients. The two bests models regarding accuracy and area under the receiver-operator characteristic (AUROC) were kept and called Model 1 (M1) and Model 2 (M2). Sensitivity, specificity, Youden index and accuracy were calculated.

8. External evaluation

The performances based on the internal test set were then compared to those on another test set. This new dataset was composed of 120 patients, whose elbow radiographs (n=262) were performed on another x-ray device (General Electrics, US) and in another time period (July to December, 2014), thus considered as an external test set and labelled by the same radiologists in the same way as the other datasets. The performances of models M1 and M2 were also measured on this dataset.

9. Evaluation of radiologists

Eight radiologists were included in this analysis: 4 radiology residents and 4 seniors radiologists specialized in musculoskeletal imaging. Radiologists had an experience in trauma radiology ranging from 6 months to 10 years.

All radiologists analysed the 120 studies of the external test set during two sessions. During the first session, radiologists interpreted all radiographs without the help of AI, then using M1 for half of the readers, and M2 for the other half of readers. After a wash-out time of two months, the second session underwent the same design (reading without AI then with AI) but model allocation was inverted. Allocation of Model 1 or 2 during the first session was distributed randomly between readers.

For each model, Radiologist interpreted x-rays without, then with the algorithm results, blinded from the other readers, and without time limitation during the session. The de-identified examinations were interpreted on their usual workstations, without access to previous or follow-up examinations to avoid follow-up bias. Radiologists were asked to first classify the study in being either normal or pathological, based on all the available x-rays for each study, prior to getting AI results, then to provide their final opinion after getting AI results.

Intra-reader agreement was measured between the two sessions for each radiologist. The sensitivity, specificity, accuracy and Youden index were calculated for each radiologist, before and after the use of AI.

10. Statistical analysis

Analyses were performed using Prism 9 software (GraphPad, La Jolla, CA). Quantitative data were reported as mean±standard deviation (SD). Qualitative data were reported as raw number and percentage (%). The significance threshold was set at $p<0.05$. To compare radiologists' performances before and after the use of AI, Wilcoxon matched-pairs signed ranked test was used for Sensitivity, Specificity, Accuracy and Youden index analysis.

IV. RESULTS

1. Internal evaluation

On the 168 patients from the internal test set, 78 had at least one pathological finding, while 90 studies were free of post-traumatic findings.

Two models stood out. The best model, M1, used cropped 512x512 images and showed the higher AUROC (0.983) and the best compromise between sensitivity (0.935) and specificity (0.978), with an accuracy of 95.83%. The second-best model, M2, used 512x512 uncropped images and presented an accuracy of 90.48%, an AUROC of 0.975, a sensitivity higher than M1 (0.974) but a lower specificity (0.844). Performances of the models on the internal test set are reported in table 4 and their receiver-operator curves (ROC) are showed in figure 2.

	Model 1 (M1)	**Model 2 (M2)**
AUROC	0.983	0.975
Accuracy	0.958	0.905
Sensitivity	0.935	0.974
Specificity	0.978	0.844
Youden index	0.913	0.818

Table 4: Evaluation of the two best models on the internal test set.

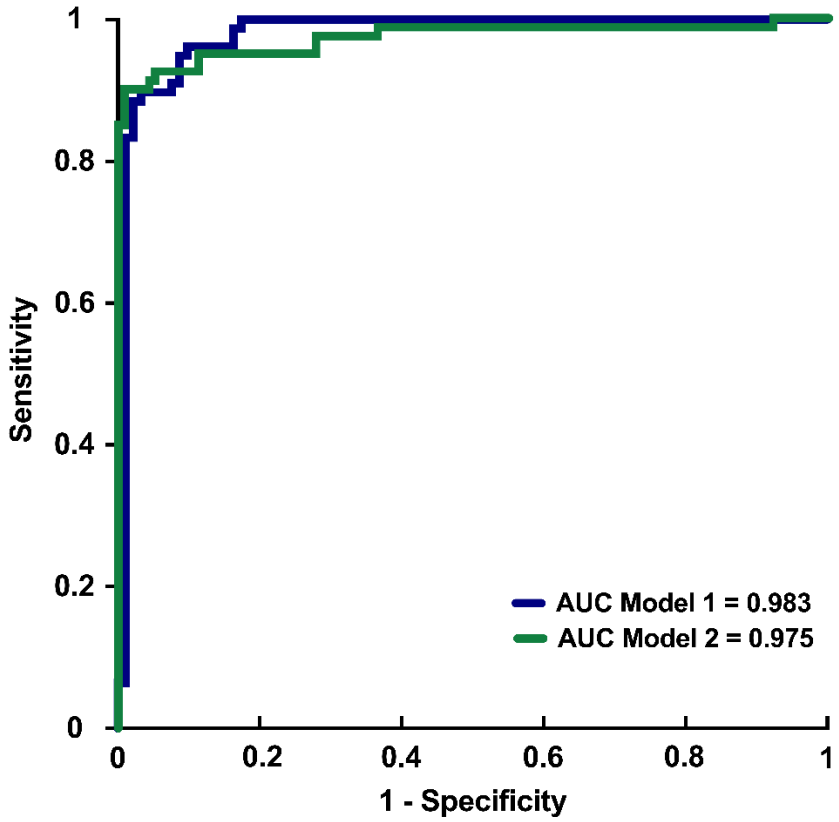

Figure 2: Receiver-operator characteristic (ROC) curves of the two best models. AUC: Area Under the Curve.

2. External evaluation of model

The external test set was composed of 61 patients with normal examinations, and 59 patients with at least one pathological finding on x-ray. The performances of M1 and M2 on this dataset are reported in Table 5.

	Model 1 (M1)	Model 2 (M2)
AUROC	0.916	0.793
Accuracy	0.825	0.692
Sensitivity	0.847	0.915
Specificity	0.803	0.475
Youden index	0.648	0.389

Table 5: Evaluation of M1 and M2 on the external test set.

While M1 showed a moderate change un AUROC values, from 0.983 to 0.916 (-0.067) M2 showed a significant drop in AUROC between internal and external test sets (0.975 to 0.793; -0.182), mostly due to a significant drop in specificity (0.844 to 0.475, -0.369).

3. Interaction with radiologists

Intra-reader agreement of examinations without AI between both sessions was excellent: 0.92 +/-0.021 for all radiologists (n=8), 0.92 +/-0.02 for radiology residents (n=4) and 0.91 +/-0.02 for senior radiologists (n=4).

Radiologists' performances without and with AI-models are reported in table 6, and represented in figure 3.

	Before M1	After M1	p	Before M2	After M2	p
Sensitivity	0.82	0.88	0.02(*)	0.83	0.85	0.06
Specificity	0.89	0.89	0.38	0.86	0.83	0.03(*)
Accuracy	0.86	0.88	0.047(*)	0.85	0.84	0.75
Youden	0.71	0.76	0.04(*)	0.69	0.69	0.94

Table 6: Radiologists' sensitivity, specificity, accuracy and Youden index before and after the use of model 1 (M1) or model 2 (M2).

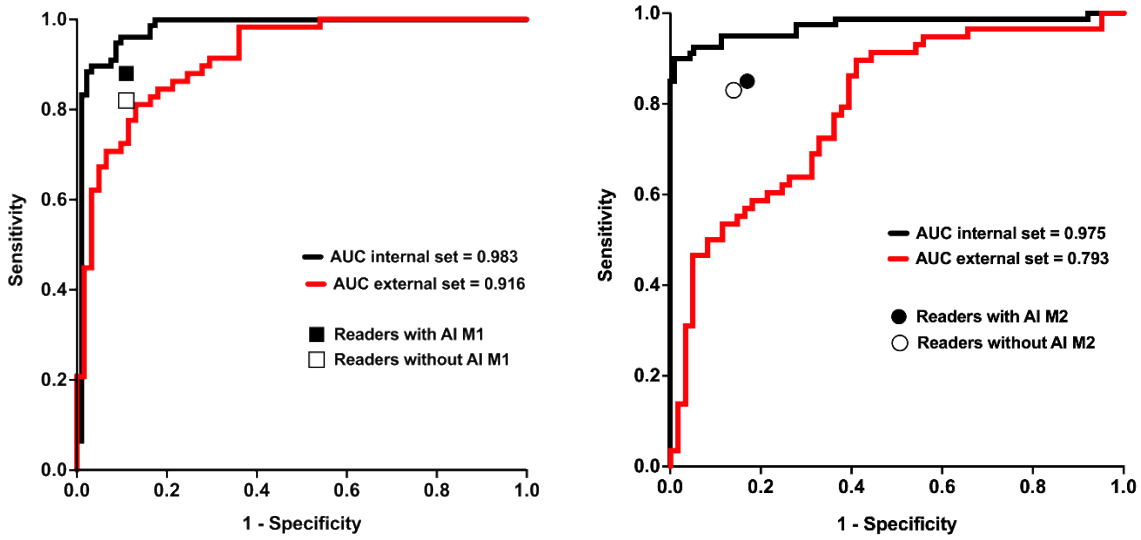

Figure 3: Variation of radiologists' Sensitivity and Specificity with the use of AI models (M1 or M2), plotted adjacent to ROC curves of both models (on the internal test set in black and on the external test set in red).

As showed in table 6 and figure 3, model 1 significantly improved radiologists' sensitivity (p=0.016), accuracy (p=0.047) and Youden index (p=0.039), while model 2 significantly decreased specificity of readers (p=0.031).

As shown in figure 3, although the initial performance of radiologists is superior to the models on the external test set, reader performances still improved significantly with the help of model 1, while model 2 did not improve the performances of readers and even significantly reduced their specificity.

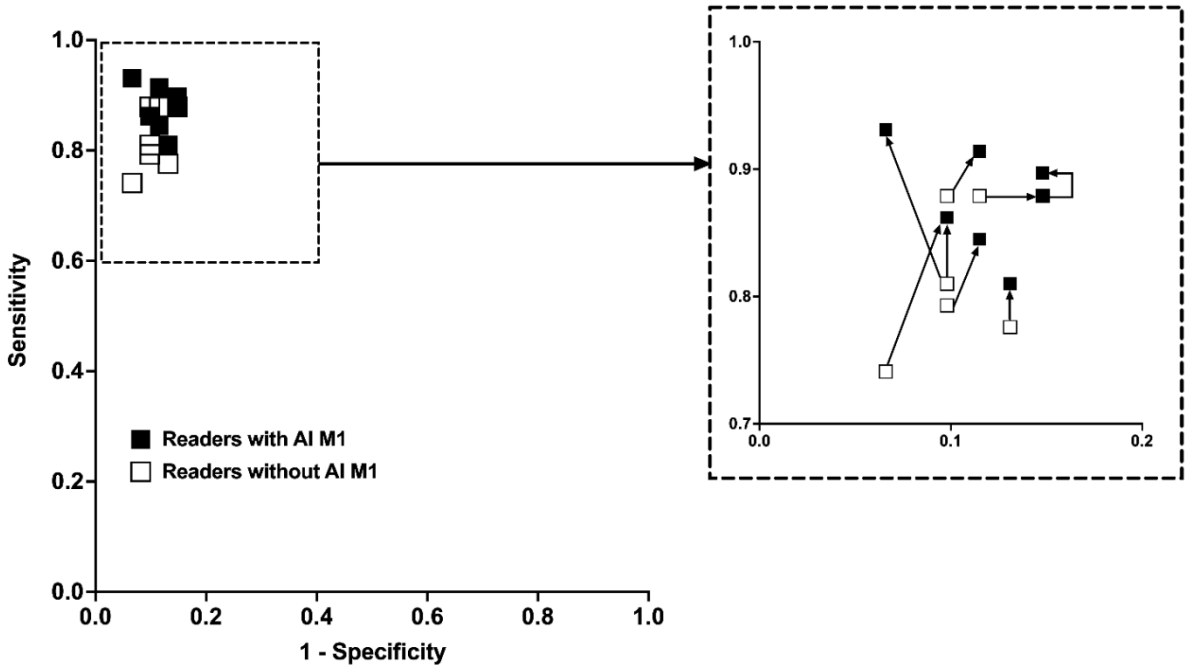

Figure 4 : All readers without then with the use of M1.

Concerning model 1, the majority of readers (n=7) increased their sensitivity but to the detriment of a slight drop in specificity for half of them (n=4). One radiologist lowered his specificity (from 0.89 to 0.85) without changing his sensitivity (0.88), therefore displaying lower performances with the help of the model.

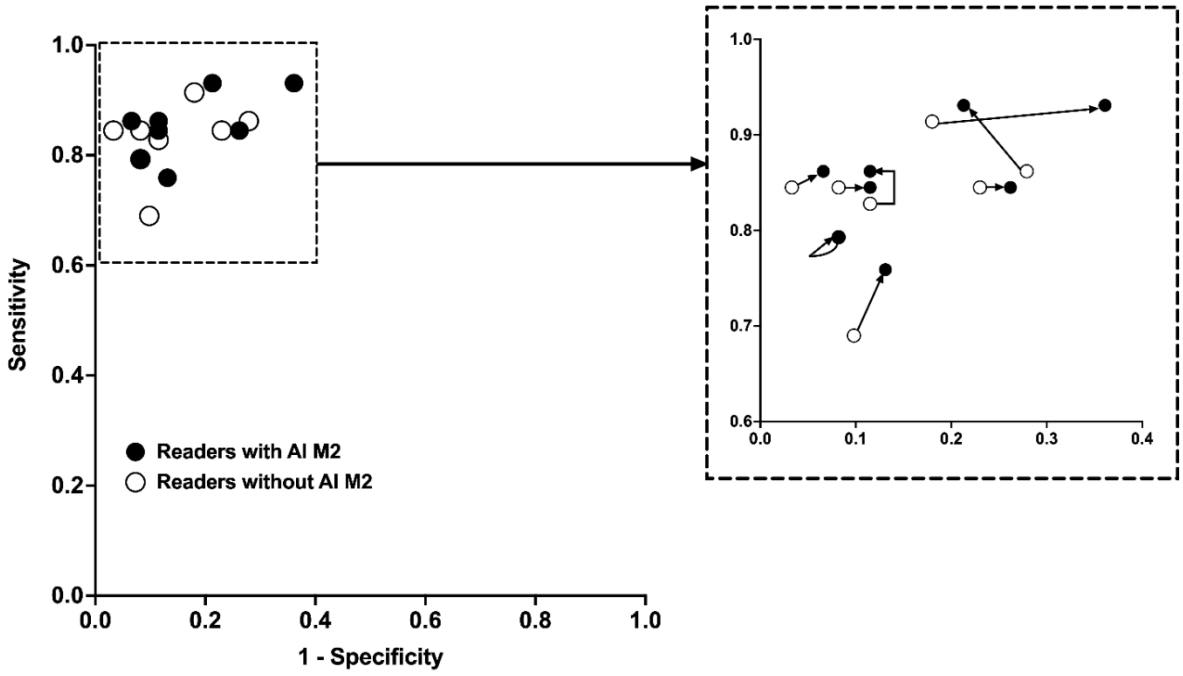

Figure 5: All readers without then with the use of M2.

Concerning model 2, more than half of radiologists (n=6) reduced their specificity, while the other (n=2) displayed no change in specificity. No reader had an improved specificity with the use of M2.

Regarding practical changes in radiologists' interpretation, figure 6 summarizes the mean number of erratic changes (*i.e.* radiologists being correct before AI but incorrect after AI) and correct changes (i.e. radiologists being incorrect before AI but correct after AI), for each model. On a radiologists' scale, M1 led to a significantly higher number of correct changes, 4.4 +/- 3.7 as compared to 2.4 +/- 2.8 for M2 (p=0.02).

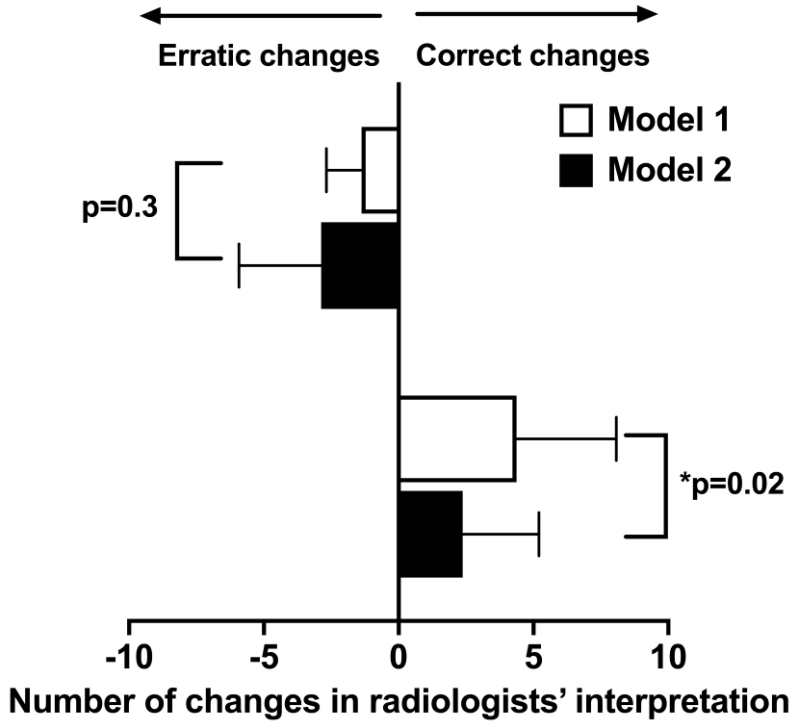

Figure 6: Correct or erratic changes in radiologist's interpretation after the use of AI, for each model.

Finally, when considering the balance of positive changes for each radiologist (number of correct changes – number of erratic changes / total number of cases), M1 displayed a positive balance of +2.2% +/- 2.7% and M2 a significantly inferior balance of -0.42% +/-1.4 (p=0.047).

4. Visualisation tool using GradCAM

As a result of all this, GradCAMs were analysed in order to check concordance of our results and ensure correct viewing of our algorithm.

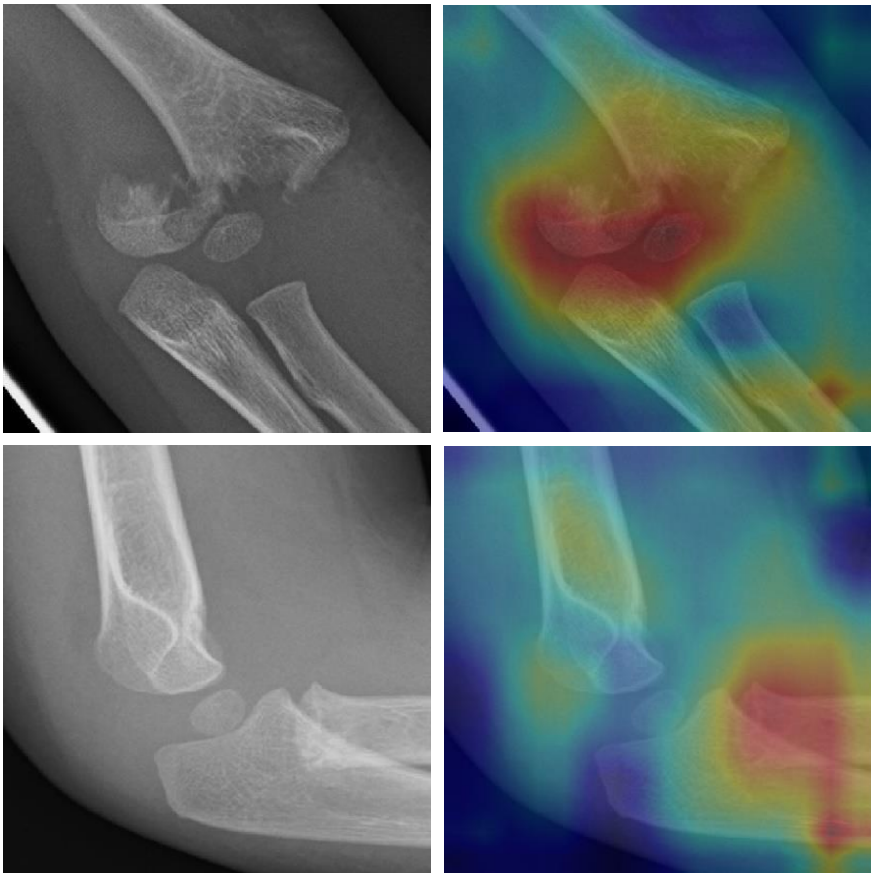

Figure 7: Two examples of true-positive GradCAMs

Figure 8: Example of false-positive case: supracondylar process which was inadequately considered as a fracture by the model.

V. DISCUSSION

This study showed that artificial intelligence using deep-convolutional neural networks can detect post-traumatic injury on elbow pediatric x-rays. AUROC values on the internal test set were high for both M1 and M2 (0.983 and 0.975 respectively), which is better than a lot of artificial intelligence publications on conventional X-rays.

Both models showed a drop in AUROC on the external test set (0.916 and 0.793 respectively), which is consistent with a tendency of DCNN-based models to overfit on internal test sets (58).However, the higher drop of M2 on the external test set raised concerns about its potential generalizability and showed that two models that display close AUROC values on internal data can undergo significantly different changes when exposed to another dataset, and that the magnitude of these changes is difficult to foresee.

Radiologists' performances were located below the internal performances of both M1 and M2 *in silico*. However, even without the help of AI, radiologists' performances were superior to those of M1 and M2 on external test sets, which shows that comparison between human and the performances algorithms on the sole internal test set should be avoided, since they tend to overclaim the inner performances of the algorithms.

Although the initial performances of radiologists were superior to those of both models on the external test set, their performances still improved significantly with the help of model 1. On the opposite, model 2 did not improve the performance of readers or even it significantly reduces their specificity. Model 2, which was supposedly the most sensitive model based on internal test set values, actually misled radiologists in their interpretation. These findings are crucial because they show that the actual impact a model can have on humans is difficult to precisely appreciate *a priori*. The

performances of M1 being slightly inferior to humans on the external test set could have implied that M1 cannot actually improve their performances, while it actually did. On the other hand, the high sensitivity of M2 *in silico* could have implied that the model would at some point increase readers' performances, while it in fact misled them more often.

Moreover, consequences of both algorithms on the radiologists' decision were measured (*i.e.* changes in individual interpretation after the use of AI). When considering practical impact on a population of patients of the use of AI by a radiologist, the key question would be the actual balance between the number of cases where changes in interpretation would indeed benefit the patient (correct changes) and changes that could impair the patient (erratic changes). Our results showed that there was a significant difference in the final benefit, since the use of M1 would results in an average gain of +2.2% in correct changes, while the use of M2 would result in a negative balance (-0.42% of correct changes).

There are some limitations of our study. First of all, though pretrained on ImageNet, our model was developed on a dataset of relatively limited size. Indeed, many algorithms focusing on conventional X-rays rely on larger datasets. Nevertheless, few (if none) have focused to the specific condition of elbow trauma in children, due to the lower availability of such data as compared to those in adults. Hence, though relatively small, our dataset is to our knowledge the largest regarding this specific question. To compensate for the size of the dataset, conventional techniques of data augmentation were performed, but are weaker to prevent overfitting as compared to fresh data, which can partly explain the changes observed between internal and external test sets. Secondly, this study was monocentric and showed that results can be variable when

exposing algorithms on an external test set. This stresses out the urge for multicentric trials in artificial intelligence. Finally, the number of readers in this study (though higher than in most published studies (59,60)) was limited (n=8), which did not enable to display differences between junior and senior readers. Further studies with more readers of different profiles are needed to confirm these results and better understand the relations between algorithms outputs and human performances.

VI. CONCLUSION

End-to-end development of a DCNN model to assess post-traumatic injuries on elbow x-ray in children was feasible, and showed that models with close metrics *in silico* can lead radiologists to either improve (M1) or lower (M2) their performances in clinical settings, underlining the need for precise clinical evaluation of AI-based tools.

Conflict of interest:

None

Funding source:

This project was funded by the CPER (Contrat de Plan Etat - Région; Région Hauts de France), involving: 1) Lille University, 2) Lille University Hospital Center and 3) INRIA Villeneuve d'Ascq.

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Larousse É. Encyclopédie Larousse en ligne - intelligence artificielle [Internet]. [cité 4 févr 2021]. Disponible sur: https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

2. Turing AM. Computing Machinery and Intelligence. Creat Comput. janv 1980;6(1):44-53.

3. Samuel AL. Programming Computers to Play Games. In: Alt FL, éditeur. Advances in Computers [Internet]. Elsevier; 1960 [cité 4 févr 2021]. p. 165-92. Disponible sur: http://www.sciencedirect.com/science/article/pii/S0065245808606087

4. Deo RC. Machine Learning in Medicine. Circulation. 17 nov 2015;132(20):1920-30.

5. Introduction au Reinforcement Learning [Internet]. [cité 25 avr 2021]. Disponible sur: https://fr.mathworks.com/discovery/reinforcement-learning.html

6. Caron M, Bojanowski P, Joulin A, Douze M. Deep Clustering for Unsupervised Learning of Visual Features. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, éditeurs. Computer Vision – ECCV 2018 [Internet]. Cham: Springer International Publishing; 2018 [cité 8 févr 2021]. p. 139-56. (Lecture Notes in Computer Science; vol. 11218). Disponible sur: http://link.springer.com/10.1007/978-3-030-01264-9_9

7. Li Y. Deep Reinforcement Learning: An Overview. ArXiv170107274 Cs [Internet]. 25 nov 2018 [cité 8 févr 2021]; Disponible sur: http://arxiv.org/abs/1701.07274

8. Tertre M. Intelligence artificielle : Comprendre le Deep et le Machine Learning [Internet]. Club de Mediapart. [cité 25 avr 2021]. Disponible sur: https://blogs.mediapart.fr/marc-tertre/blog/130318/intelligence-artificielle-comprendre-le-deep-et-le-machine-learning

9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. mai 2015;521(7553):436-44.

10. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. août 2018;18(8):500-10.

11. ImageNet [Internet]. [cité 23 avr 2021]. Disponible sur: http://www.image-net.org/index

12. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 1 déc 2015;115(3):211-52.

13. Russakovsky O, Deng J, Huang Z, Berg AC, Fei-Fei L. Detecting Avocados to Zucchinis: What Have We Done, and Where Are We Going? In: 2013 IEEE International Conference on Computer Vision [Internet]. Sydney, Australia: IEEE; 2013 [cité 23 avr 2021]. p. 2064-71. Disponible sur: http://ieeexplore.ieee.org/document/6751367/

14. Fathers of the Deep Learning revolution receive 2018 ACM A.M. Turing Award [Internet]. [cité 23 avr 2021]. Disponible sur: https://www.acm.org/media-center/2019/march/turing-award-2018

15. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol. 23 nov 2020;

16. Ranjan E, Paul S, Kapoor S, Kar A, Sethuraman R, Sheet D. Jointly Learning Convolutional Representations to Compress Radiological Images and Classify Thoracic Diseases in the Compressed Domain. In: Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing [Internet]. New York, NY, USA: Association for Computing Machinery; 2018 [cité 5 févr 2021]. p. 1-8. (ICVGIP 2018). Disponible sur: https://doi.org/10.1145/3293353.3293408

17. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. J Magn Reson Imaging JMRI. avr 2019;49(4):939-54.

18. Wacker J, Ladeira M, Nascimento JEV. Transfer Learning for Brain Tumor Segmentation. ArXiv191212452 Cs Eess [Internet]. 26 nov 2020 [cité 23 avr 2021]; Disponible sur: http://arxiv.org/abs/1912.12452

19. Paul R, Hawkins S, Balagurunathan Y, Schabath M, Gillies R, Hall L, et al. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival among Patients with Lung Adenocarcinoma. Tomography. 1 déc 2016;2(4):388-95.

20. Albahli S, Alhassan F, Albattah W, Khan RU. Handwritten Digit Recognition: Hyperparameters-Based Analysis. Appl Sci. janv 2020;10(17):5988.

21. Leen TK, Dietterich TG, Tresp V. Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference. MIT Press; 2001. 1136 p.

22. What is Underfitting? [Internet]. 2021 [cité 22 avr 2021]. Disponible sur: https://www.ibm.com/cloud/learn/underfitting

23. Sabottke CF, Spieler BM. The Effect of Image Resolution on Deep Learning in Radiography. Radiol Artif Intell. 1 janv 2020;2(1):e190015.

24. Maurice B. Data augmentation [Internet]. Deeply Learning. 2018 [cité 27 mai 2021]. Disponible sur: https://deeplylearning.fr/cours-theoriques-intelligence-artificielle/data-augmentation/

25. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp. 24 oct 2018;2(1):35.

26. AI-Index-2018-Annual-Report.pdf [Internet]. [cité 22 avr 2021]. Disponible sur: https://covenworks.com/wp-content/uploads/2019/01/AI-Index-2018-Annual-Report.pdf

27. Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: What do radiologists think? Diagn Interv Imaging. 1 juin 2019;100(6):327-36.

28. Rao B, Zohrabian V, Cedeno P, Saha A, Pahade J, Davis MA. Utility of Artificial Intelligence Tool as a Prospective Radiology Peer Reviewer — Detection of Unreported Intracranial Hemorrhage. Acad Radiol. 1 janv 2021;28(1):85-93.

29. Lakhani P, Prater AB, Hutson RK, Andriole KP, Dreyer KJ, Morey J, et al. Machine Learning in Radiology: Applications Beyond Image Interpretation. J Am Coll Radiol. 1 févr 2018;15(2):350-9.

30. Kannampallil T, Smyth JM, Jones S, Payne PRO, Ma J. Cognitive plausibility in voice-based AI health counselors. Npj Digit Med. 15 mai 2020;3(1):1-4.

31. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. Npj Digit Med. 4 avr 2018;1(1):1-7.

32. Sandino CM, Cheng JY, Chen F, Mardani M, Pauly JM, Vasanawala SS. Compressed Sensing: From Research to Clinical Practice With Deep Neural Networks: Shortening Scan Times for Magnetic Resonance Imaging. IEEE Signal Process Mag. janv 2020;37(1):117-27.

33. Philips Illumeo with adaptive intelligence has been selected by University of Utah Health radiologists [Internet]. Philips. [cité 21 avr 2021]. Disponible sur: https://www.philips.com/a-w/about/news/archive/standard/news/press/2018/20181126-philips-illumeo-with-adaptive-intelligence-has-been-selected-by-university-of-utah-health-radiologists.html

34. Dey D, Commandeur F. Radiomics to identify high-risk atherosclerotic plaque from CT: the power of quantification. Circ Cardiovasc Imaging [Internet]. déc 2017 [cité 11 févr 2021];10(12). Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5731252/

35. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 22 sept 2014;5(1):4006.

36. Guly H. Diagnostic errors in an accident and emergency department. Emerg Med J EMJ. juill 2001;18(4):263-9.

37. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop. déc 2017;88(6):581-6.

38. Ebsim R, Naqvi J, Cootes TF. Automatic Detection of Wrist Fractures From Posteroanterior and Lateral Radiographs: A Deep Learning-Based Approach. In: Vrtovec T, Yao J, Zheng G, Pozo JM, éditeurs. Computational Methods and Clinical Applications in Musculoskeletal Imaging. Cham: Springer International Publishing; 2019. p. 114-25. (Lecture Notes in Computer Science).

39. Chung SW, Han SS, Lee JW, Oh K-S, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. 4 juill 2018;89(4):468-73.

40. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol. févr 2019;48(2):239-44.

41. Kitamura G, Chung CY, Moore BE. Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation. J Digit Imaging. 1 août 2019;32(4):672-7.

42. Roth HR, Wang Y, Yao J, Lu L, Burns JE, Summers RM. Deep convolutional networks for automated detection of posterior-element fractures on spine CT. In: Medical Imaging 2016: Computer-Aided Diagnosis [Internet]. International Society for Optics and Photonics; 2016 [cité 5 févr 2021]. p. 97850P. Disponible sur: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9785/97850P/Deep-convolutional-networks-for-automated-detection-of-posterior-element-fractures/10.1117/12.2217146.short

43. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. Radiol Artif Intell. janv 2019;1(1):e180001.

44. Lapègue F, Sans N. Echographie musculosquelettique. Elsevier Masson; 2014. 424 p.

45. Imagerie musculosquelettique : pathologies locorégionales [Internet]. [cité 23 mai 2021]. Disponible sur: https://www.elsevier-masson.fr/imagerie-musculosquelettique-pathologies-locoregionales-9782294721823.html

46. Bledsoe RC, Izenstark IL. Displacement of Fat Pads in Disease and Injury of the Elbow: A New Radiographic Sign. Radiology. nov 1959;73(5):717-24.

47. O'Dwyer H, O'Sullivan P, Fitzgerald D, Lee MJ, McGrath F, Logan PM. The fat pad sign following elbow trauma in adults: its usefulness and reliability in suspecting occult fracture. J Comput Assist Tomogr. août 2004;28(4):562-5.

48. Abzug JM, Herman MJ. Management of supracondylar humerus fractures in children: current concepts. J Am Acad Orthop Surg. févr 2012;20(2):69-77.

49. Mary P. La pronation douloureuse. In: Cahiers d'enseignement de la SOFCOT n° 72. p. 222-6. (Paris : Expansion Scientifique Française 2000).

50. Imagerie Musculosquelettique : Pathologies Générales [Internet]. Elsevier; 2013 [cité 5 févr 2021]. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/C20120075916

51. Hambidge SJ, Davidson AJ, Gonzales R, Steiner JF. Epidemiology of pediatric injury-related primary care office visits in the United States. Pediatrics. avr 2002;109(4):559-65.

52. Kraus R, Wessel L. The Treatment of Upper Limb Fractures in Children and Adolescents. Dtsch Ärztebl Int. déc 2010;107(51-52):903-10.

53. McGinley JC, Roach N, Hopgood BC, Kozin SH. Nondisplaced elbow fractures: A commonly occurring and difficult diagnosis. Am J Emerg Med. sept 2006;24(5):560-6.

54. Bisset GS, Crowe J. Diagnostic errors in interpretation of pediatric musculoskeletal radiographs at common injury sites. Pediatr Radiol. mai 2014;44(5):552-7.

55. Home [Internet]. [cité 22 mai 2021]. Disponible sur: https://www.gleamer.ai/

56. Duron L, Ducarouge A, Gillibert A, Lainé J, Allouche C, Cherel N, et al. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. Radiology. 4 mai 2021;203886.

57. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 19 févr 2021;4(1):31.

58. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research:: A clinical example. J Clin Epidemiol. 1 sept 2003;56(9):826-32.

59. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. juin 2019;25(6):954-61.

60. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2 janv 2020;577(7788):89-94.

AUTEUR : Nom : ROZWAG **Prénom : Clémence**

Date de soutenance : 01/07/2021

Titre de la thèse : Développement et évaluation de modèles d'intelligence artificielle pour la détection des lésions post-traumatiques du coude de l'enfant en radiographie.

Thèse - Médecine - Lille 2021

Cadre de classement : *Radiodiagnostic et Imagerie Médicale*

DES + spécialité : *DES de Radiodiagnostic et Imagerie Médicale*

Mots-clés : Radiographie traumatologique, Coude pédiatrique, Intelligence artificielle

Résumé :

CONTEXTE : Les erreurs de diagnostic sur les radiographies traumatologiques sont courantes, notamment chez l'enfant. De nombreux modèles d'intelligence artificielle (IA) ont été développés pour détecter les lésions post-traumatiques mais peu ont été validés chez l'enfant. De plus, l'impact de ces modèles en pratique clinique est rarement évalué. L'objectif était de développer un modèle d'IA capable de détecter les lésions post-traumatiques sur les radiographies du coude pédiatrique, d'évaluer ses performances puis l'impact de son utilisation par les radiologues.
MÉTHODE : 1956 radiographies pédiatriques du coude post-traumatique ont été collectées chez 935 patients âgés de 0 à 18 ans. Elles ont été réparties aléatoirement en trois jeux de données : données d'entraînement, de validation et de test interne. Les réseaux de neurones convolutifs profonds ont été formés en faisant varier les hyperparamètres (données d'entraînement et de validation) puis évalués sur les données de test interne. Les deux meilleurs modèles sélectionnés ont été évalués sur des données de test externe impliquant 120 patients, dont les radiographies ont été réalisées avec un équipement radiologique différent à une autre période. Huit radiologues ont interprété ces données sans l'aide des modèles d'IA puis avec.
RÉSULTATS : Deux modèles se sont démarqués sur les données de test interne. Le modèle 1 (M1) avait une précision de 95.8% et une AUROC de 0.98, avec le meilleur compromis Sensibilité (Se)/Spécificité (Sp) (Se=0.93/Sp=0.97). Le modèle 2 (M2) avait une précision de 90.5% et une AUROC de 0.97 avec une meilleure Se mais une Sp plus faible (Se=0.97/Sp=0.84). Sur le jeu de test externe, M1 a conservé une bonne précision de 82.5% et AUROC de 0.916 (-0.067) ainsi qu'un bon compromis Se/Sp (Se=0.84/Sp=0.80). En revanche, M2 a connu une baisse de performance avec une perte de précision jusqu'à 69.2% et d'AUROC à 0.793 (-0.182). M1 a significativement amélioré la Se du radiologue (0.82 à 0.88 ; p=0.02) et la précision (0.86 à 0.88 ; p=0.04), tandis que M2 a significativement diminué la Sp des lecteurs (0.86 à 0.83; p=0.03).
CONCLUSION : Le développement d'un modèle d'IA pour détecter les lésions post-traumatiques du coude pédiatrique en radiographie est faisable mais des modèles avec des performances proches *in silico* peuvent conduire les radiologues à améliorer (M1) ou à baisser (M2) leurs performances en pratique clinique, soulignant la nécessité d'une évaluation clinique précise des outils basés sur l'IA.

Composition du Jury :

Président : Madame le Professeur A. COTTEN

Assesseurs : Monsieur le Professeur X. DEMONDION et Monsieur le Professeur P. PREUX

Directeur de thèse : Monsieur le Docteur T. JACQUES