UNIVERSITÉ DE LILLE

FACULTÉ DE MÉDECINE HENRI WAREMBOURG
Année 2024

THÈSE POUR LE DIPLÔME D'ÉTAT
DE DOCTEUR EN MÉDECINE

Intérêt pour les urgentistes, de la pré analyse des radiographies traumatiques, par l'intelligence artificielle

Présentée et soutenue publiquement le 17 octobre 2024 à 18h
au Pôle Formation
par Ilona BELLOTTO

_____

JURY
Président :
Monsieur le Professeur Eric WIEL
Assesseurs :
Monsieur le Docteur Jean-Marie RENARD
Monsieur le Docteur Jérôme MIZON
Directeur de thèse :
Monsieur le Docteur Romain DEWILDE

_____

AVERTISSEMENT

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

ABBRÉVIATIONS

AUC : area under the curve = aire sous la courbe

CH : centre hospitalier

CHM : centre hospitalier de Maubeuge

COMU : Collège de médecine d'urgence

DL : deep learning

DPO : délégué(e) à la protection des données

DREES : Direction de la recherche, des études, de l'évaluation et des statistiques

DIU : diplôme inter-universitaire

DU : diplôme universitaire

IA : intelligence artificielle

IC : intervalle de confiance

IQR : interquartile range = intervalle interquartile

ML : machine learning

NC : non communiqué

PACS : Picture Archiving and Communications System

ROC : Receiver Operating Characteristic

Se : sensibilité

SFMU : société française de médecine d'urgence

SOFCOT : société française de chirurgie orthopédique et traumatologique

Sp : spécificité

SSA : service de santé des armées

TDM : tomodensitométrie

VPN : valeur prédictive négative

VPP : valeur prédictive positive

LISTES DES FIGURES

LISTES DES TABLEAUX

SOMMAIRE

RÉSUMÉ

Contexte : l'augmentation des passages aux urgences et le manque de médecins urgentistes et radiologues a conduit à une surcharge de travail. Aux urgences, l'interprétation des radiographies est souvent réalisée par des urgentistes non formés spécifiquement, augmentant ainsi le risque d'erreurs, notamment pour les fractures. Le développement de l'intelligence artificielle (IA), comme l'outil Milvue Suite, vise à améliorer la qualité et l'efficacité du diagnostic dans ce contexte.

Objectifs : l'objectif principal de cette étude était d'évaluer l'impact de l'utilisation de Milvue Suite sur les performances diagnostiques des urgentistes exerçant dans la lecture des radiographies traumatiques. L'étude visait à déterminer si l'IA améliore la sensibilité et la spécificité des urgentistes dans la détection des lésions traumatiques. Les objectifs secondaires étaient d'évaluer l'effet de l'IA sur le temps de lecture des radiographies et la comparaison des performances entre internes et séniors, avec et sans IA.

Matériels et Méthodes : étude rétrospective monocentrique ayant inclus 49 cas de radiographies traumatiques d'adultes, associées à des TDM comme référence diagnostique. Les radiographies étaient annotées en deux phases : une première sans IA, puis une deuxième avec l'aide de Milvue Suite. 30 lecteurs (24 internes et 6 séniors) ont participé à cette étude. Les performances diagnostiques (Se, Sp, VPP et VPN) et le temps de lecture ont été comparés entre les deux phases.

Résultats : L'IA Milvue Suite a amélioré la Se des lecteurs, avec une augmentation globale de 68 % à 88 %. L'amélioration des performances diagnostiques était plus marquée chez les séniors, avec un gain de sensibilité de 23%, contre 18% chez les internes. En revanche, il n'a pas été possible d'interpréter la Sp et VPN en raison du faible nombre de cas négatifs dans l'échantillon. De plus, l'IA a permis de réduire significativement le temps médian de lecture, passant de 88,95s sans IA à 48,08s avec IA, représentant une réduction de 40,87s.

Conclusion : L'IA Milvue Suite a amélioré la Se et réduit le temps de lecture des radiographies traumatiques par les urgentistes, à la fois chez les séniors et les internes. Cependant, le faible effectif de cas constitue une limite de l'étude. L'IA apparaît comme un outil prometteur pour soutenir les urgentistes.

INTELLIGENCE ARTIFICIELLE (IA) : DÉFINITION – HISTOIRE – TYPES D'IA – IA EN SANTÉ – MILVUE

- **Définition**

1) Définition de l'intelligence artificielle

L'intelligence artificielle est un ensemble de théories et de techniques informatiques et mathématiques capable de simuler le raisonnement humain.

2) Différence entre intelligence humaine et artificielle

L'intelligence est la capacité d'un système vivant à comprendre, interpréter, apprendre et s'adapter aux changements. Elle est incarnée, elle est indissociable du corps. (1)
L'IA n'est pas incarnée et se limite toujours à un domaine défini. Elle se base sur l'apprentissage et elle a donc besoin de faire des statistiques sur un nombre important de données, contrairement à l'intelligence humaine, qui peut faire des déductions pertinentes à partir de quelques exemples. (1)

- **Histoire de l'IA**

Le concept d'IA est lancé en 1950 par le mathématicien britannique Alan M. Turing, lorsqu'il met au point un « test d'imitation » capable de déterminer si une machine peut simuler les réponses d'un être humain. (2)
L'IA devient une discipline scientifique à part entière, reconnue comme domaine de recherche en 1956, lors de la conférence de Dartmouth, organisée par John McCarthy et Marvin Minsky. Au cours de cette conférence a été présenté le « *Logic Theorist »* (écrit en 1955 par A. Newell, H. Simon, C. Shaw), un programme informatique conçu pour reproduire les compétences de résolution de problèmes d'un être humain. (3)
En 1958, J. McCarthy invente le langage Lisp au MIT *(Massachusetts Institute of Technology),* un des langages de programmation les plus simples en termes de syntaxe. (4)

Jusqu'en 1974, l'IA prospère grâce aux financements et aux recherches prometteuses, mais les objectifs finaux mettant du temps à être atteint, les financements sont peu à peu coupés et les projets sur l'IA s'essoufflent. C'est le premier hiver de l'IA et prendra fin au début des années 1980.

L'intelligence artificielle est relancée dans les années 1980-1990, d'une part grâce aux travaux de John Hopfield, David Rumelhart et Yann Le Cun qui popularisent le « deep learning », permettant aux ordinateurs d'apprendre à s'appuyant sur l'expérience (5,6). Et d'autre part, grâce à Edward Feigenbaum, qui introduit les systèmes experts, imitant le processus de décision d'un expert humain.

Le gouvernement japonais finance jusqu'en 1992 les systèmes experts et d'autres projets en lien avec l'IA, à hauteur de 400 millions de dollars (7). Mais encore une fois, les objectifs peinent à être atteint, les financements cessent, faisant tomber l'IA dans l'oubli du grand public. L'IA connait ici sont 2ème hiver, et prendra fin vers les années 2010.

Malgré l'absence de financement et de médiatisation, la communauté scientifique poursuit ces investigations sur le domaine et notamment via le jeu d'échecs. Notons que c'est à partir des années 1960 que les « Chess programs » commencent à se développer. Bien que la vitesse et la puissance de calcul de ces algorithmes s'améliorent (la loi de Moore énoncée en 1965 par George Moore, estime que la vitesse et la puissance des ordinateurs doublent tous les 18 mois (8)), ils restent systématiquement surpassés par les maitres de la discipline, comme en témoigne la défaite en 1989 du programme « Deep Thought » d'IBM, sur le champion du monde Garri Kasparov.

Les recherchent s'affinent et les programmes s'améliorent, ce qui permet à la machine de prendre sa revanche sur l'être humain des décennies plus tard, en mai 1997, où le programme « Deeper Blue » l'emporte sur Kasparov et en 2016 où AlphaGo de Google Deepmind bat le champion de jeu de go, Lee Sedol.

Bien que la loi Moore tende à ralentir[i], la capacité à collecter et stocker[ii] des données s'est améliorée. Ajouter à l'accroissement exponentiel de la puissance du numérique, le volume de données numériques créé a explosé. On estime que ce volume est de l'ordre de 40% en moyenne par an sur 5 ans et qu'il s'est multiplié par 30 en 10 ans (9) (Figure 1). En 2021, il est estimé qu'1 minute sur Internet génère en moyenne 198 millions d'e-mails, 695 000 stories partagées sur Instagram et 500 heures de contenus mis en ligne du YouTube (10).

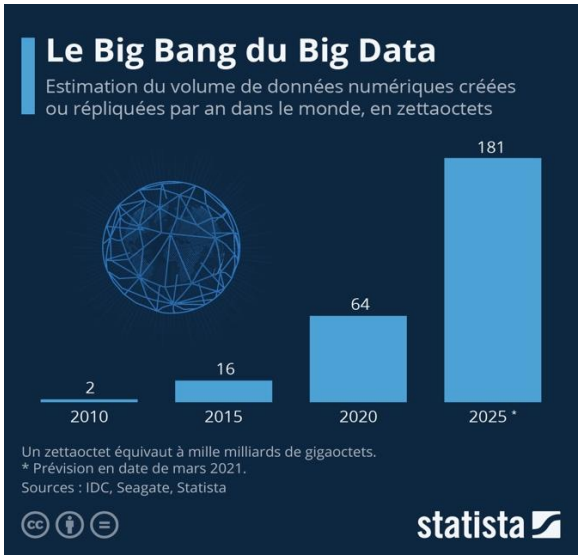
L'explosion quantitative des données numériques a donné naissance au concept de « big data » dans les années 2000. Un exemple qui illustre parfaitement ce concept, est ImageNet, crée en 2009 par la chercheuse Fei-Fei Li (11). C'est un jeu de données composé de plus de

[i] Par le fait que la taille des puces électroniques se heurte à une limite physique, qui est celle de l'atome

[ii] Notamment via les data center ou entrepôts de données et via la gravure des puces électroniques en 2nm

14 millions d'images étiquetées[i], réparties en millier de catégories. Ce projet a joué un rôle crucial dans les progrès de l'intelligence artificielle et notamment dans l'apprentissage automatique, la vision par ordinateur et la reconnaissance d'images ou d'objets. En raison de son énorme volume de données, de la variété des types d'images, de l'évolution continue des données, ImageNet est le parfait exemple du Big Data et son utilité pour entrainer les algorithmes d'intelligence artificielle.

Figure 1: Le Big Bang du Big Data (source : statitsta.com; Tristan Gaudiaut; 19/10/2021)

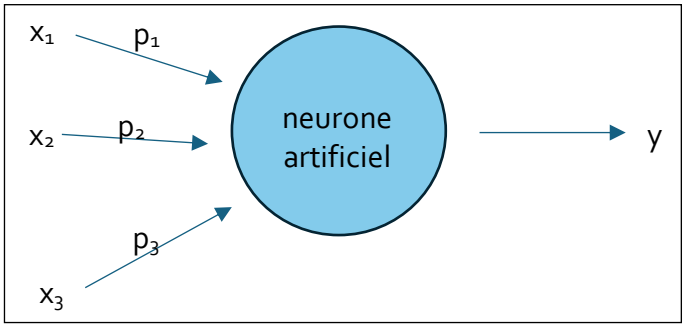


[i] 3,2 millions lors de sa création en 2009

- **Machine Learning (ML) et Deep Learning (DL)**

L'apprentissage automatique ou « Machine Learning » est un ensemble d'opérations informatiques permettant aux ordinateurs d'apprendre automatiquement en étudiant des exemples. Arthur Samuel, est le 1er, en 1959, après la création d'un programme pour IBM, à utiliser ce terme. Selon les données disponibles, l'apprentissage peut être supervisé, non supervisé ou par renforcement. Le ML se base sur le concept du Perceptron, crée en 1957 par le psychologue Frank Rosenblatt (12). Le perceptron est un neurone artificiel modélisé par une fonction mathématique mettant en relation des entrées « x » et des sorties « y ». Ce neurone mime le fonctionnement d'un neurone biologique, et donc, l'apprentissage humain : recevoir un ou des signaux des neurones voisins avec ce dont il est connecté et en fonction de ces signaux reçus, il peut ou pas envoyer à son tour des signaux aux autres neurones. Le neurone artificiel va recevoir plusieurs entrées « x » et en faire la somme. Chaque entrée « x » est affecté à un coefficient qu'on appelle un « poids ». Si la somme pondérée est supérieure à un certain seuil, le neurone délivre une sortie « y » égale à 1, si la somme est inférieure, la sortie « y » sera égale à 0 (Figure 2).

Figure 2 - Représentation d'un neurone artificiel ou Perceptron

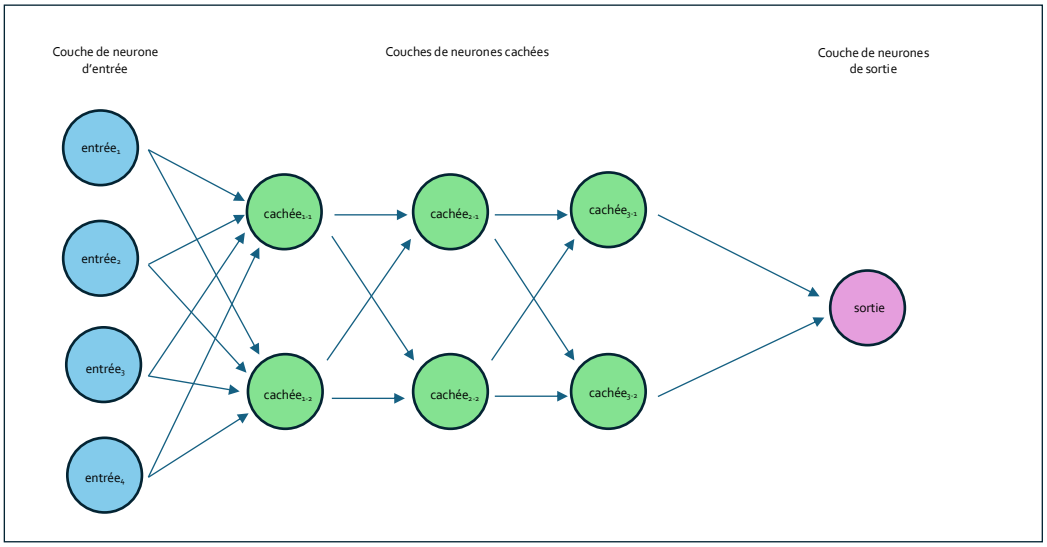


** p correspond au poids et est représenté par les flèches à gauche du cercle*

Lorsque les données entrantes dans l'algorithme sont nombreuses et/ou le lien entre ces données entrantes et les prédictions sortantes sont complexes, les algorithmes d'IA utilisent des « réseaux de neurones profonds ». C'est ce que l'on appelle apprentissage profond ou « Deep Learning », qui est une sous-catégorie du ML. Le DL utilise un empilement de ces neurones artificiels pour former un réseau de neurones artificiels profonds. Il y a une couche de neurone entrante ou « input layer », des couches de neurones cachées ou « hidden layers » et enfin, une couche de neurone sortante, « output layer ». Les neurones de chaque couche sont connectés aux couches adjacentes par des connexions, appelées encore « poids » ou « weights» (Figure 3).

La fonction mathématique qui en découle sera d'autant plus complexe que le réseau neuronal contient de couches cachées.

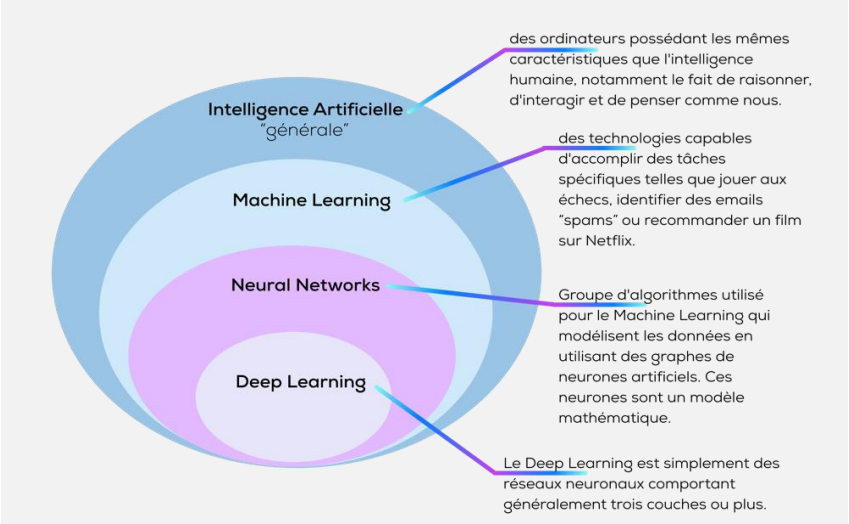
Figure 3 - Représentation simplifiée d'un réseau de neurone profond



Bien que le Deep Learning soit un sous-ensemble du Machine Learning, il existe des différences :

- **Volume d'informations** et l'**intervention humaine** : le DL nécessite un volume important de données pour s'entrainer, mais apprend par son propre environnement et de ses propres erreurs. A l'inverse, le ML, nécessite un volume de données plus faible mais l'intervention de l'humain est nécessaire pour apprendre et corriger les erreurs : un humain doit labelliser et caractériser les données.
- **Le temps d'entrainement** : la faible quantité d'informations rend l'entrainement du ML plus rapide alors que le temps d'entrainement des algorithmes de DL est bien plus long
- **Le niveau de précision** : les algorithmes de DL est plus précis que ceux de ML car traitent des données beaucoup plus complexes et plus nombreuses
- **L'approche** : là où les algorithmes de DL considèrent un problème dans son entièreté, ceux de ML cherchent à séparer les données puis les recombiner pour proposer une solution

Figure 4 - Diagramme de Venn de l'intelligence artificielle (source : site web Allison.com ; 24/04/2024)



- **<u>L'IA dans le domaine de la santé, en radiologie et aux urgences</u>**

Un des 1ers programme d'intelligence artificielle utilisé en médecine, a vu le jour dans les années 1977, à l'université de Stanford, *Californie*. Edward Shortliffe et son maitre de thèse Bruce G. Buchanan, développent le système expert *« MYCIN »* permettant d'identifier des bactéries responsables d'infections graves, de recommander une antibiothérapie dont les posologies sont adaptées au poids du patient (13). En avril 2015, le géant de l'informatique IBM lance sa filiale Watson Health. Cette dernière était une initiative du groupe américain qui visait à utiliser l'intelligence artificielle, en particulier leur superordinateur Watson, pour révolutionner le domaine de la santé. L'idée était que Watson, qui avait déjà gagné au jeu télévisé « Jeopardy! », pourrait aider les médecins à diagnostiquer des maladies, proposer des traitements, et analyser des données médicales complexes. Cependant le programme a été un échec parce que les attentes étaient trop élevées voire irréalistes, la technologie n'était pas prête, les données entrantes étaient de mauvaise qualité et le produit final n'était pas assez utile pour les professionnels de la santé. Les recommandations de Watson n'étaient pas toujours fiables, pas toujours compréhensibles voire étaient erronées, ce qui a fragilisé la confiance des professionnels de santé avec le programme américain. IBM a finalement dû vendre la majeure partie de Watson Health en 2022, reconnaissant que le projet n'avait pas réussi à transformer le domaine de la santé comme prévu.
Les données entrantes dans les algorithmes d'IA ne se doivent pas uniquement d'être nombreuses mais d'être de qualité et structurées.

De nombreux progrès ont été fait dans le domaine de l'IA en santé grâce au Big Data, à l'informatisation des établissements de santé, à la numérisation des images et des dossiers patients. La pratique de la médecine implique des heures passées à devenir un expert dans la reconnaissance de modèles et la résolution de problèmes de haut niveau. Le processus

d'apprentissage est long. Les radiologues et les urgentistes par exemple, au cours de leur carrière, examinerons des milliers de clichés radiologiques. Ils passent de nombreuses années à construire une base de données de références mentale, et ils l'affinent au fil de leur carrière. Une des premières spécialités à avoir suscitée de l'intérêt pour l'IA, est l'imagerie médicale, via l'interprétation radiographique. Des algorithmes d'aide à l'interprétation des radiographies ont vu le jour, à l'instar de Milvue Suite™, développé par Milvue. Grâce au « Computer Vision » ou « vision par ordinateur » utilisant la reconnaissance d'images et les réseaux de neurones récurrents *(popularisé par Yann Le Cun en 1990 via la lecture optique des chèques)*, le processus d'apprentissage a été accéléré. Là où le médecin devra voir des milliers de patients étalés sur des dizaines d'années pour avoir une confiance forte en son diagnostic, l'algorithme, lui, pourra agréger une quantité importante de données d'entrainement en quelques semaines seulement.

- **Milvue**

Milvue est une entreprise française créée en 2018, ayant pour objectif de développer des logiciels d'IA dans le domaine de l'imagerie médicale. Son équipe a développé Milvue Suite, solution d'IA de détection et de mesure, et notamment TechCare Alert pour les radiographies d'urgences. Milvue Suite, permet de détecter 8 pathologies, à savoir : épanchement articulaire du coude, épanchement pleural, fracture du gril costal, fracture du squelette appendiculaire, luxation articulaire, nodule et opacité pulmonaire, pneumothorax. (14)
Chaque cliché radiographique est annoté d'une boite, englobant les zones lésionnelles et les zones de doute. Les clichés sont donc triés en 3 catégories : présence, doute ou absence de lésion.

Une récente étude de V. Bousson en juillet 2023, a comparé 3 solutions d'IA française dans l'aide à la lecture des radiographies aux urgences, dont Milvue Suite. Bien que les 3 logiciels eussent une sensibilité (Se) élevée et voisines, la solution de Milvue s'est avérée être celle dont l'exactitude était la plus élevée. Aussi, Milvue Suite a démontré être la solution qui produisait le moins de faux positif, et des performances quasi identiques peu importante la région anatomique analysée, à l'inverse des algorithmes concurrents. (15)

En s'avérant être un outil performant, cette étude nous rassure quant à l'utilisation de Milvue suite dans notre pratique courant aux urgences du CH de Maubeuge.

INTRODUCTION

Depuis 1996, l'activité des urgences en France ne cesse de croitre : le nombre de passages a été multiplié par 2 entre 1996 et 2019[i] (Figure 5). En 2022, les urgences de France ont pris en charge 21,6 millions de passages, soit 6,2% de plus qu'en 2021. Ces chiffres sont similaires aux urgences du CH de Maubeuge (+ 6,9% entre 2021 et 2022, annexe 2). On constate aussi une augmentation du nombre de passages par structure d'urgences, dû à la diminution du nombre de structure et, en parallèle, de l'augmentation du nombre de passage (16). A cette augmentation de la charge de travail aux urgences, s'ajoute, un manque de médecin, avec des lignes de gardes non pourvues[ii] (17) et aussi, un nombre de radiologues hospitaliers et libéraux qui tant à évoluer en sens inverse depuis 2012, avec, une diminution des premiers cités, et une augmentation des seconds[iii] (18). Pourtant, 40% des consultations aux urgences donnent lieu à une imagerie (19), dans 71% des cas il s'agit d'un radiographie standard. Les résultats sont similaires aux urgences du Maubeuge et ce depuis 2021 (annexe 2). En effet, le principal motif de consultation aux urgences est le traumatisme (20). Le nombre de demandes d'examens d'imageries aux urgences ne cesse donc de croitre et il en devient difficile pour les radiologues d'interpréter en temps réel cette quantité accrues de clichés. C'est pourquoi dans de nombreux centres, c'est l'urgentiste qui se charge de cette tâche, bien qu'il n'y soit pas formé. A l'heure actuelle, en France, la maquette du Diplôme d'études spécialisées de médecine d'urgence DESMU, n'inclut pas de stage ni de formation avancée obligatoire en imagerie (21) ce qui pourrait causer des erreurs de diagnostics, sachant que la majeure partie des erreurs médicales faites aux urgences sont les fractures (22). Notons aussi que 27% des patients se présentent aux urgences la nuit (20h-8h) et 27% le week-end (23), plages horaires où d'une part l'interprétation des imageries par un radiologues est plus difficile à obtenir et d'autre part, la fatigue accumulée par le travail de nuit peut induire des erreurs (24).

C'est donc, dans ce contexte d'accumulation de la fatigue par augmentation de la charge de travail, de jour comme de nuit, d'augmentation du risque d'erreur diagnostique, de manque de médecins urgentistes et de radiologues, que l'intelligence artificielle se développe au sein des structures d'urgences et de radiologie. Certains centres hospitaliers se munissent d'algorithmes d'intelligence artificielle d'aide à la lecture radiographique, comme c'est le cas depuis 2019, dans le Nord, aux urgences du CH de Maubeuge. Ce CH utilise actuellement aux urgences, Milvue Suite, une IA distribuée par la société française Milvue. T. Jacques et al. a

[i] Hormis en 2020, où les urgences ont connu une baisse inédite du nombre de passage (- 17,7%), en lien avec la pandémie COVID-19.

[ii] L'enquête nationale menée par la DREES, « Urgences 2023 », a montré que 19% des points d'accueil d'urgences ont eu une ligne de travail non pourvue la semaine suivant l'enquête.

[iii] Une diminution de 2% entre 2012 et 2023 pour les radiologues hospitaliers et une augmentation de 3,5% pour les radiologues libéraux sur la même période.

montré que l'IA permet aux radiologues d'améliorer significativement leur Se et le VPN pour la détection des fractures du poignet et de la main en radiographie, sans affecter ni leur Sp ni la VPP (25). A. Parpaleix et al, ont démontré récemment, que ces logiciels d'IA ont une performance diagnostique élevée, une Se similaire à celle des urgentistes et peuvent réduire les erreurs diagnostiques sur les imageries musculosquelettiques et thoraciques (26).

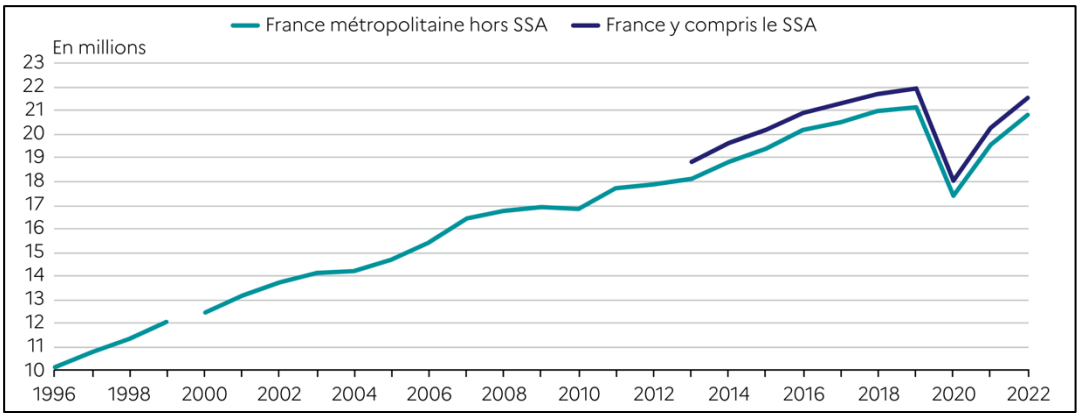
N'étant pas spécialiste dans l'interprétation d'images, une aide cognitive comme les logiciels d'IA, pourraient permettre aux urgentistes d'argumenter un faisceau de preuve avec un degré de certitude élevé. Il nous est donc apparu opportun de voir l'impact de l'outil Milvue Suite sur des praticiens juniors et séniors des urgences du CH de Maubeuge, alors qu'il est désormais utilisé en routine dans ce service.

Notre hypothèse de départ est que, l'utilisation de l'algorithme de Milvue par les internes de médecine d'urgences et les séniors des urgences du CHM, dans l'aide à la lecture des radiographies traumatologiques, améliore leurs performances de lecture ou du moins, ne les détériore pas.

Notre objectif principal sera de chercher si les performances diagnostiques des séniors et internes, exerçant aux urgences du CHM, à la lecture des radiographies traumatologiques, sont améliorés avant et après utilisation de l'algorithme d'intelligence artificielle Milvue Suite, par rapport à la base de données référentielle *(appelée ground-truth)* basée sur la tomodensitométrie.

Nos objectifs secondaires seront d'abord de montrer que les performances de lecture de ces mêmes radiographies par les mêmes parties sus citées ne sont pas détériorées après aide de ce même logiciel d'intelligence artificielle. Puis, de déterminer quel est l'impact de Milvue Suite sur la vitesse de lecture des lecteurs.

Figure 5 - Évolution du nombre de passage annuels aux urgences depuis 1996 *(source : DREES, SAE 1996-2022, traitements DREES)*



Note 1 : la modification du questionnaire relatif aux urgences et la référence aux articles définissant l'activité de soins autorisée à compter de l'enquête SAE 2000 introduisent une rupture de série entre 1999 et 2000
Note 2 : France métropolitaine hors SSA de 1996 à 2022, France (incluant Saint-Martin et Saint-Barthélemy) y compris le SSA de 2013 à 2022*
** SSA : service de santé des armées*

MATÉRIELS ET MÉTHODES

- **Dataset**

1) Création du jeu de données initial ou « dataset »

La première étape de notre étude rétrospective, monocentrique et observationnelle, a été la construction du jeu de données. Ce dernier correspondait à un ensemble de radiographies de patients âgés d'au moins 18 ans pris en charge pour traumatisme ostéoarticulaire, des membres supérieurs et/ou inférieurs aux urgences du centre hospitalier de Maubeuge (CHM, Maubeuge, France) ainsi que les tomodensitométries (TDM ou scanner) qui s'y associaient. Ces TDM ont été fait en 2ème intention, après radiographie, lorsqu'il y a eu un doute sur le diagnostic ou une incohérence clinico-radiologique. Les TDM ont été faits dans l'intérêt du patient et de sa prise en charge, les patients n'ont pas été surirradiés dans le but de notre étude.

Tout d'abord, nous avons recherché de manière rétrospective via le PACS, sur la période février 2022 à juin 2024 les clichés radiographiques d'intérêt.

- Les critères d'inclusion ont été : adulte d'au moins 18 ans au moment de la radiographie, victime d'un traumatisme de la main et/ou du poignet et/ou du coude et/ou de la cheville et/ou du pied et/ou de la hanche, ayant bénéficié, dans le cadre du traumatisme pour lequel il se présentait aux urgences, d'une radiographie puis d'un scanner. Les 2 examens devaient être fait aux urgences du CHM, le même jour.
- Les critères d'exclusion ont été : les patients de moins de 18 ans au moment de leur passage aux urgences, les victimes d'un traumatisme du rachis dans son ensemble, cérébral, grill costal et/ou thoracique, abdominal, d'une autre articulation non sus citée. Les patients dont le dossier médical mentionnait des antécédents d'ostéoporose et de cancers osseux ont été exclus.

Pour se faire, nous avons d'abord interrogé le PACS grâce à l'outil de filtrage. Nous avons entré successivement, les mots clés suivant, « main », « poignet », « coude », « hanche », « cheville » et « pied » et pour chaque zone anatomique recherchée, sélectionné la modalité d'imagerie « scanner ». Nous avons ensuite effectué la même recherche en sélectionnant cette fois la modalité « radiographie ».
Puis nous avons fait un « matching » manuel, entre date de radiographie et de scanner, nom, prénom, date de naissance et zone anatomique : un même patient, majeur, (vérifier via le nom, prénom et date de naissance) ayant eu à la même date une radiographie et un scanner de la

même zone anatomique, était incluable dans notre dataset. A cette étape, nous avons vérifié, via le logiciel Sillage, dans le dossier médical de chaque patient potentiellement incluable, l'absence d'ostéoporose ou de cancer osseux mentionné dans les antécédents. Notons que les patients dont les antécédents n'étaient pas renseignés sur le logiciel Sillage, ont tout de même été inclus.

2) Anonymisation des patients

La deuxième étape a été de télécharger sur notre serveur local et sécurisé, pour chaque cas inclus, les radiographies et le scanner correspondant. Au moment du téléchargement via le PACS, nous avons d'abord dé-identifié chaque cliché. Ensuite, nous avons créé des dossiers propres à chaque cas, que nous avons renommer via un code alphanumérique *(un code par cas inclus dans l'étude)*. Cette méthode nous a permis d'anonymiser nos données tout en gardant une table de correspondance. Cette étape a été validée par la Déléguée à la Protection des Données (DPO), du CHM.

3) Données cliniques récoltées

En parallèle, nous avons recherché via Sillage, pour chaque patient adulte inclus dans l'étude, les éléments suivants :

- l'âge au moment du traumatisme, sans récolter la date de naissance
- la date de la radiographie et du scanner
- le sexe du patient
- les antécédents s'ils étaient décrits *(la mention NC a été utilisée si les ATCD n'avaient pas été signalé dans le dossier médical)*
- la zone anatomique traumatisée et son côté droit, gauche ou bilatéral
- le contexte du traumatisme s'il était décrit *(la mention NC a été utilisée si les ATCD n'avaient pas été signalé dans le dossier médical)*
- l'examen clinique s'il était décrit *(la mention NC a été utilisée si les ATCD n'avaient pas été signalé dans le dossier médical)*
- le diagnostic principal retenu sur le scanner

Ces données ont été classées dans un tableur Excel et associées pour chaque cas, à son code alphanumérique précédemment créé.
L'ensemble des données contenues dans le tableau Excel étaient anonymes. A ce stade de l'étude, il ne nous était plus possible d'identifier les patients inclus dans le dataset.

- **Transfert des données**

1) Plateforme BlueFiles

La solution française BlueFiles permet d'effectuer des transferts de données sécurisés, en leur appliquant un chiffrement de bout en bout. L'ensemble des services de transmission de données par BlueFiles sont accessibles à partir de son site internet.
BlueFiles dispose du Visa de Sécurité de l'ANSSI, l'identifiant ainsi comme une solution fiable en matière de cybersécurité. L'hébergeur du service BlueFiles est certifié SecNumCloud, il s'agit du plus haut niveau d'engagement en matière de sécurité. L'hébergeur de BlueFiles est certifié HDS (Hébergement des Données de Santé), gage de confiance en termes de sécurité pour le traitement des données de santé à caractère personnel. BlueFiles est d'une part en conformité avec les recommandations de la CNIL et d'autre part respecte le RGPD via le « Privacy by Design », le droit à l'oubli, la portabilité et la traçabilité.

2) Transfert des données à Milvue

Pour transférer nos dossiers alphanumérisés à Milvue, nous avons utilisés la plateforme BlueFiles. Nous avons simplement envoyé des emails contenant notre dataset *(divisé en dossier par zone anatomique)* ainsi que le fichier Excel, à Milvue, en passant par BlueFiles. Le chiffrement des données de bout en bout a été assuré comme expliqué précédemment par BlueFiles. Nous rappelons que Milvue n'a reçu aucune donnée permettant d'identifier précisément les patients inclus dans l'étude.

3) Utilisation des données par Milvue

Les radiographies dé-identifiées ont été intégrées par Milvue sur la plateforme d'annotation Encord. Les scanners dé-identifiés associés à chaque radiographie a servi de ground-truth, pour l'annotation par les urgentistes des clichés radiographiques. Seul Milvue a eu la possibilité de lire les comptes rendus anonymisés. Notons que les fichiers n'ayant pas pu être téléchargés sur la plateforme Encord, ont été exclus du protocole. Les données cliniques recueillies dans le tableur Excel, ont permis lors de l'analyse des données par Milvue, d'apporter des précisions démographiques sur le dataset et sur les lecteurs à notre étude.

L'intégralité du dataset et du fichier Excel sera supprimé par Milvue, au lendemain de la soutenance de cette thèse.

- **Annotation des radiographies**

1) Les lecteurs

Les radiographies dé-identifiées ont été interprétées par un groupe de lecteurs.
Les lecteurs étaient, internes, docteur juniors (DJ) ou praticiens hospitaliers ayant fini leur formation. Les internes et DJ faisaient tous partis du DESMU à Lille. Les praticiens pratiquaient tous dans le service d'urgences du CH de Maubeuge.
Pour l'analyse des résultats nous avons réuni praticiens et docteurs juniors dans un même groupe nommé « séniors ».

2) L'interprétation

L'interprétation des radiographies s'est faite sur la plateforme d'annotation Encord. Nous avons fourni à chaque lecteur, via e-mail, un compte utilisateur pour avoir accès à la plateforme. Pour chaque phase, ils ont aussi reçu via e-mail, une pièce jointe, expliquant la procédure d'annotation correspondante. Une phase d'entrainement a été demandée aux lecteurs, afin de prendre en main la plateforme et les outils d'annotation.

L'interprétation s'est faite en 2 phases :

- Phase 1 : une période d'annotation du 22 au 27 juillet 2024 jours sans l'aide de l'algorithme Milvue Suite. Pour chaque cas, les lecteurs devaient d'abord analyser toutes les incidences proposées en positionnant des annotations sur la ou les zones considérées comme pathologique(s), ou aucune annotation en cas de radiographies considérées comme normales. Les annotations étaient enregistrées automatiquement à la clôture de chaque cas.
- Période de wash out de 4 semaines du 27 juillet au 26 aout 2024
- Phase 2 : une période d'annotation du 26 aout au 31 aout 2024, avec l'aide de l'algorithme cette fois. De la même manière que sur la 1ère période de lecture, les lecteurs devaient d'abord annoter chaque cliché radiographique avant de pouvoir activer l'IA. Ces annotations ont été enregistrées automatiquement. A l'aide des résultats proposés par l'algorithme, les lecteurs pouvaient ensuite ajouter, retirer ou modifier les annotations en fonction de leur analyse définitive. Les annotations définitives étaient enregistrées automatiquement à la clôture de chaque cas.

Sur les 2 périodes de lecture, le temps de lecture pour chacun des lecteurs a été récolté.

Les radiographies ont été présentées dans un ordre aléatoire pour chaque phase, qui était différent pour chacun des lecteurs.
Pour chaque lecteur, la sensibilité (Se), spécificité (Sp), valeur prédictive positive (VPP), valeur prédictive négative (VPN), et le temps de lecture sans puis avec utilisation de l'IA, ont été calculés, en se basant sur le *ground-truth* issu du scanner.
Les mesures de performances sus citées ont été calculées à la fois à l'échelle des cas et à l'échelle des lésions individuelles, pour mieux évaluer les modifications de détection de chaque fracture. Précisons qu'à l'échelle de la lésion, la Sp et la VPN n'ont pas pu être calculés car à cette échelle on part du principe qu'il n'y a que des cas positifs.
Une analyse en sous-groupe en fonction de l'ancienneté des lecteurs (interne vs séniors) et en fonction de la localisation de la fracture a été réalisés afin d'analyser les performances de lecture. L'aire sous la courbe (AUC) de la courbe ROC de l'algorithme a été calculée ainsi que la Se, Sp, VPP et VPN à l'échelle des cas et des lésions. Les définitions de Se, Sp, VPP, VPN et AUC sont données dans l'annexe 1. Les lecteurs n'ayant pas fini ou pas commencé les annotations des radiographies des deux phases ont été exclus des analyses statistiques.

- **<u>Aspect règlementaire et éthique</u>**

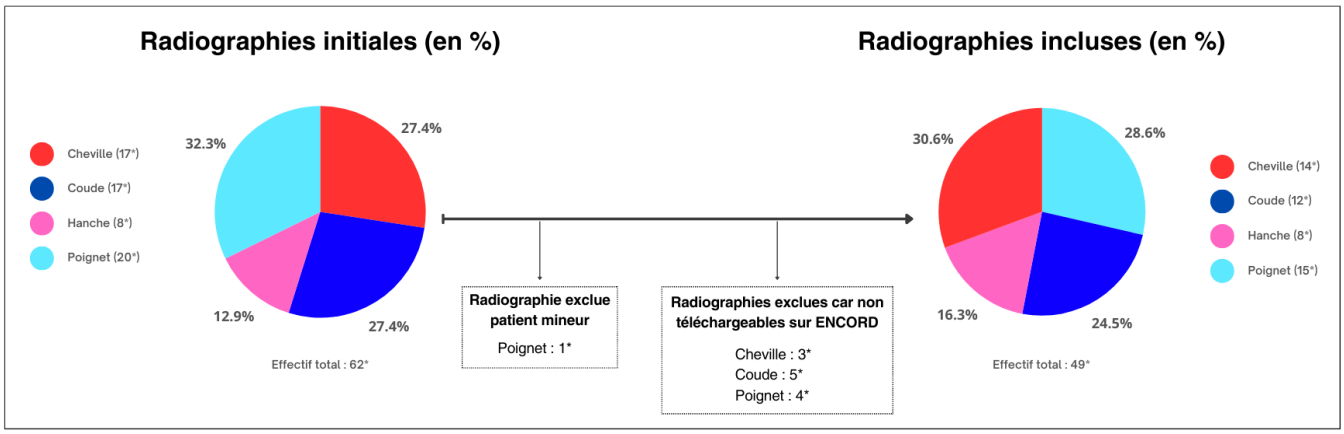
Tous les accords réglementaires et éthiques nécessaires à cette étude ont été demandés ou assurés. Une convention de collaboration pour la recherche médicale a été signée par le CHM et la société Milvue. L'article 11.1 de cette convention mentionne le respect du RGPD. Un registre des traitements a été rempli par les auteurs de cette thèse, la société Milvue et L. Lemoine, DPO au Centre Hospitalier de Maubeuge.

RÉSULTATS

- **Le jeu de données**

Notre jeu de données initial était constitué de 62 clichés radiographiques[i] : 17 cas de cheville, 17 cas de coude, 8 cas de hanche et 20 cas de poignet. 12 cas n'ont pas pu être téléchargés sur la plateforme ENCORD (3 cas de cheville, 5 cas de coude et 4 cas de poignet) et 1 cas de poignet était un patient mineur. Nous avons donc exclu de l'analyse ces 13 cas (Figure 6). Les données démographiques du jeu de données initial (avant exclusion) sont données en annexe 3.

Figure 6 : Jeu de données initial et d'analyse réparti par zone anatomique



** Effectif en nombre*

Notre jeu de données inclus pour l'analyse est constitué de 49 clichés radiographiques[ii]. 8 cas de hanche, 14 cas de cheville, 12 cas de coude et 15 cas de poignet. 18 patients étaient des hommes et 31 patients étaient des femmes. Le sex ratio de 0,6 était en faveur du groupe femme (Tableau 1). La médiane d'âge de notre population était de 66 ans, avec une médiane d'âge plus faible chez les hommes que chez les femmes (40 vs 69 ans respectivement).
On remarque que la moyenne d'âge était plus basse que la médiane, 57,8 ans contre 66 ans respectivement ce qui signifie que la répartition des âges était asymétrique avec une queue à droite et donc une part importante de personnes âgées. Les âges extrêmes de notre population sont 19 ans et 92 ans.
Les sous-groupes poignet et cheville avaient un âge médian inférieur (55 et 58,5 ans de médiane d'âge respectivement) aux sous-groupes hanche et coude (76,5 vs 67,5 ans respectivement). Le sous-groupe hanche avait l'âge médian le plus élevé (Tableau 1).

[i] Notre jeu de données initial comprenait aussi le TDM correspondant à chaque radiographie (donc 62 TDM)

[ii] Notre jeu de données d'analyse comprenait aussi le TDM correspondant à chaque radiographie (donc 49 TDM). Chaque radiographie exclue a vu son TDM correspondant être exclure de l'analyse

Notons aussi, que le sous-groupe d'âge < 65 ans était en faveur des hommes avec un sex-ratio de 1,2 ; alors que le sous-groupe ≥ 65 ans était en faveur des femmes avec un sex-ratio de 0,3 (Tableau 1).

Les caractéristiques du ground-truth (GT) du jeu de données initial, avant exclusion, est donné en annexe 3. Dans le jeu de données d'analyse[i], le GT était positif pour 45 cas (91,8%), c'est-à-dire qu'au moins une lésion de type fracture et/ou luxation était présente, et négatif pour 4 cas (Tableau 2). On constate donc que notre jeu de données était déséquilibré, en faveur des lésions. Cela suggère que les patients adressés pour des TDM avaient une suspicion clinique forte de lésion sur la radiographie, ce qui est une source de biais de sélection.

Concernant la répartition par sexe des cas positifs et négatifs, le GT était positif pour l'ensemble des patients masculins (18 cas) (100%). Sur un total de 31 femmes, le GT était positif pour 27 cas (87,1%) et négatif pour 4 cas (Tableau 2).

La totalité des cas de hanche (n=8) et de cheville (n=14) avaient un GT positif. Le GT était également positif pour 11 cas de coude (91,7% ; 12 cas au total) et pour 12 cas de poignet (80,0% ; 15 cas au total) (Tableau 2).

Concernant la latéralité des zones anatomiques lésées, il y avait 25 cas de traumatisme à droite (51%), 21 cas à gauche (42,9%) et 3 cas bilatéraux (6,1%). 2 traumatismes de hanche étaient à droite, 5 à gauche et 1 des 2 cotés (25%, 62,5% et 12,5% respectivement). Pour les traumatismes de cheville, 9 étaient à droite (64,3%) et 5 à gauche (35,7%). Pour les traumatismes de coude, 7 étaient à droite (58,3%) et 5 à gauche (41,7%). Pour les traumatismes de poignet, 7 étaient à droite, 6 à gauche et 2 des deux côtés (46,7%, 40%, 13,3% respectivement) (Tableau 2).

[i] Patients inclus dans l'analyse statistique des résultats

Tableau 1 : Caractéristiques de la population incluse

	Age *(années)*						Effectif *(nb)*	Sex ratio H/F
	Moyenne	Médiane	1er quartile	3ème quartile	Min	Max	Total	
Population globale	57,8	66	37	75	19	92	49	0,6
Homme	47,3	40	31,3	66,8	20	87	18	
Femme	63,8	69	60	76	19	92	31	
≥ 65 ans	75,1	73	68	79	65	92	27	0,3
< 65 ans	36,5	35,5	24	41	19	63	22	1,2
Hanche	72,1	76,5	66,8	88,3	24	90	8	1
Cheville	53,8	59,5	35,3	72,5	19	92	14	0,3
Coude	59,6	67,5	54,8	70,8	20	87	12	0,7
Poignet	52,3	55	38,0	70,5	19	78	15	0,7
Note 1 : nb signifie nombre								

Tableau 2 : Caractéristiques du ground-truth du jeu de données d'analyse

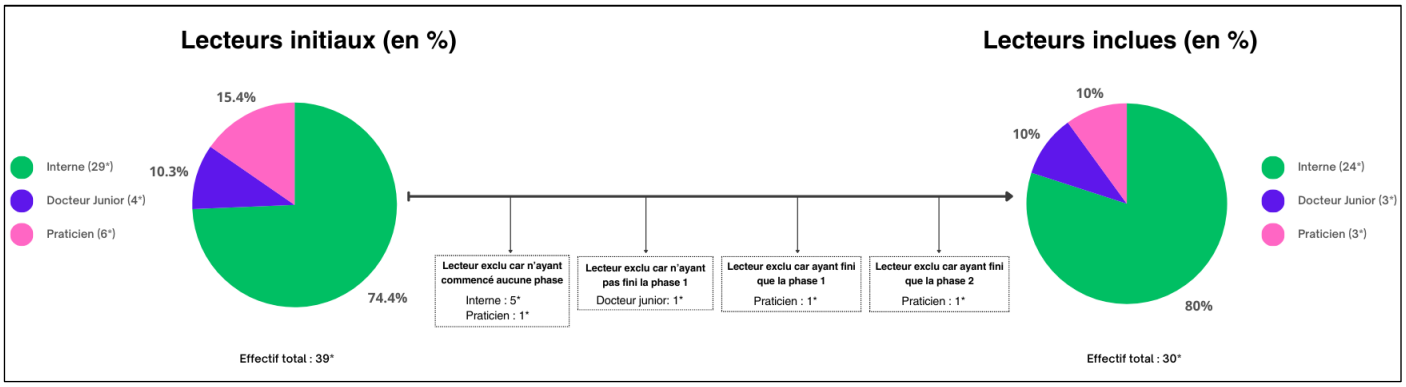
	TDM *(nb)*		Latéralité *(nb)*		
	Positif	Négatif	Droit	Gauche	Droit et Gauche
Population générale	45	4	25	21	3
Homme	18	0	13	5	0
Femme	28	3	12	16	3
Hanche	8	0	2	5	1
Cheville	14	0	9	5	0
Coude	11	1	7	5	0
Poignet	12	3	7	6	2
Note 1 : nb signifie nombre					

- **Les lecteurs**

Initialement 39 lecteurs se sont portés volontaires pour participer aux annotations des radiographies. 5 lecteurs n'ont participé à aucune des 2 phases d'annotations, nous les avons donc exclus de l'analyse. Parmi eux on comptait 4 internes et 1 praticien. Lors de la première phase d'annotation, 8 séniors dont 4 praticiens et 4 docteurs juniors, et 24 internes ont annotés des images. Lors de la deuxième phase d'annotation, 24 internes, et 8 séniors dont 4 praticiens et 4 docteurs juniors ont annotés des images (Figure 7).

Un des praticiens n'a participé qu'à la phase 1, et un autre praticien n'a participé qu'à la phase 2 d'annotation. Nous avons donc exclu ces 2 lecteurs car ils n'ont pas participé aux 2 phases du protocole. Un docteur junior a été exclu de l'analyse car il n'a pas fait la totalité des annotations lors de la phase 1 (bien qu'il ait fait la totalité des annotations de la phase 2) (Figure 7).

Figure 7 : Diagramme de flux du groupe de lecteurs

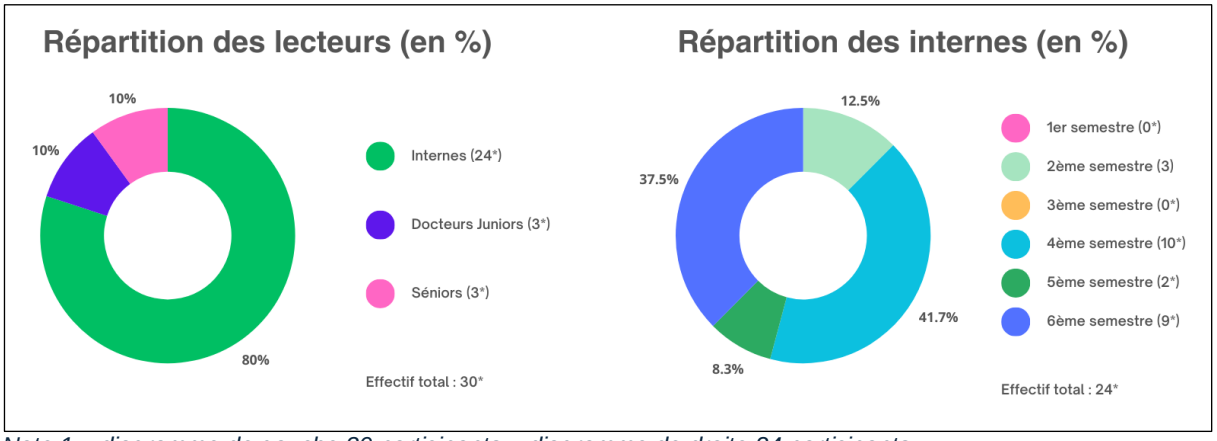


** Effectif en nombre*

Au total, 6 séniors (3 DJ et 3 praticiens) et 24 internes, soit 30 lecteurs, ont participé à la totalité du protocole et ont été inclus dans l'analyse des résultats. 80% des lecteurs étaient internes et 20% étaient séniors dont 10% de docteurs juniors. 8,3% des internes étaient dans le 5ème semestre de leur cursus, 12,5% dans le 2ème semestre, 41,7% dans le 4ème semestre et 37,5% dans le 6ème semestre. Aucun néo interne[i] ou interne de 3ème semestre n'a participé à cette étude (Figure 8). La médiane d'année d'expérience du groupe « sénior » est de 2,8 ans.

[i] Interne de 1er semestre

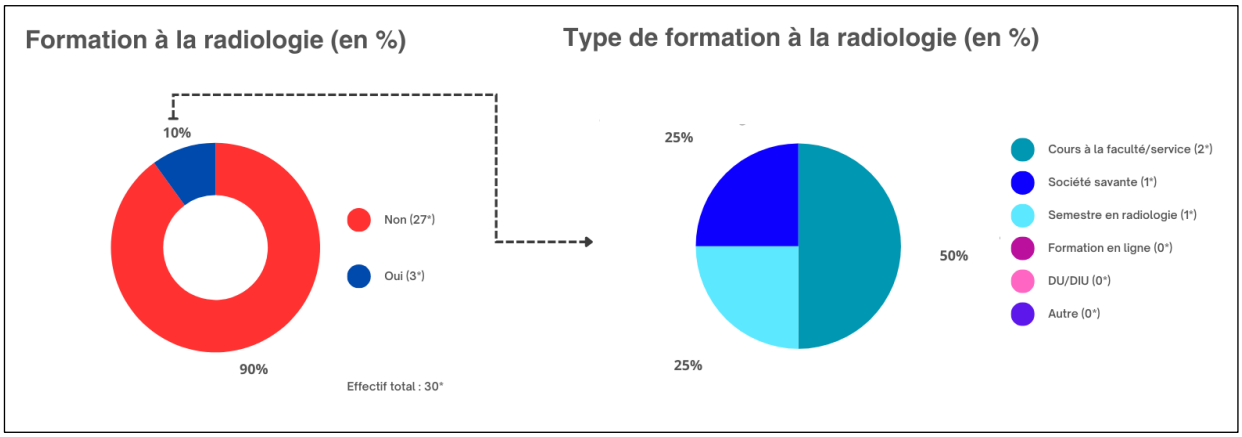
Figure 8 : Statut des lecteurs et des internes inclus dans l'analyse



Note 1 : diagramme de gauche 30 participants – diagramme de droite 24 participants
*Note 2 : * représente l'effectif en nombre*

90% des lecteurs ont déclarés ne pas avoir eu de formation à la radiologie durant leur internat ou leur carrière en tant que praticien. 10% ont dit avoir eu une formation à la radiologie. Parmi eux, le type de formation correspondait pour 25% à un semestre en service de radiologie, 25% à une formation via une société savante (type COMU, SFMU, SFR) et 50% lors de cours à la faculté de médecine de Lille ou au sein d'un service (Figure 9).

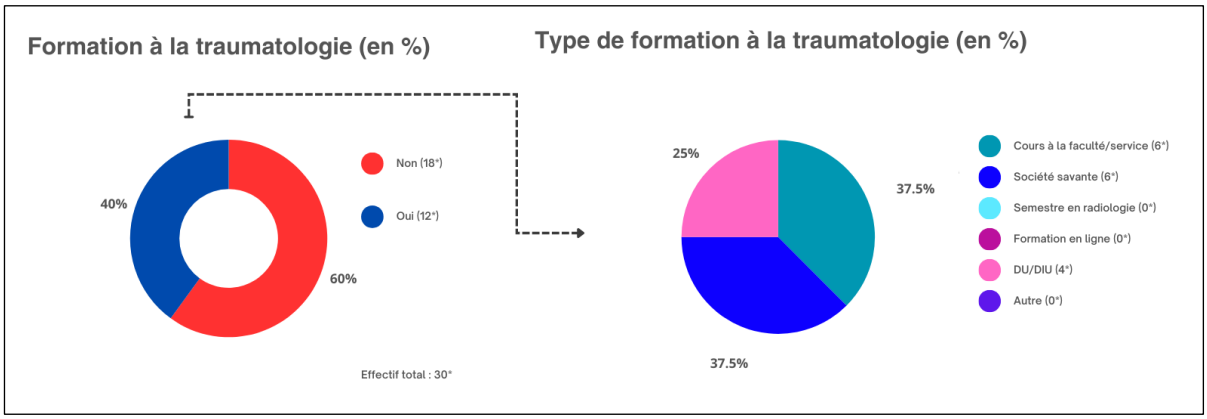
Figure 9 : Formation et type de formation à la radiologie



Note 1 : diagramme de gauche 30 participants – diagramme de droite 4 participants
*Note 2 : * représente l'effectif en nombre*

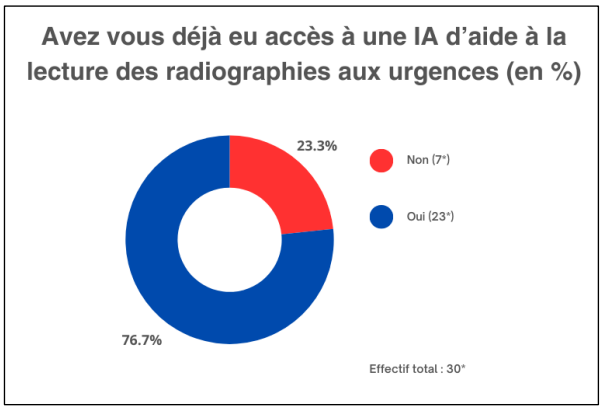
De la même façon, nous nous sommes intéressés à la formation en traumatologie de nos lecteurs. 60% ont déclarés avoir eu une formation en traumatologie au cours de leur internat ou en post internat. Parmi eux, 37,5% ont eu une formation sous la forme de cours à la faculté de médecine de Lille ou au sein d'un service, 37,5% ont eu une formation via une société savante (type COMU, SFMU, SOFCOT) et 25% ont eu une formation via un DU ou DIU (Figure 10). Nous avons aussi questionné nos lecteurs sur le fait d'avoir déjà eu accès à un IA d'aide à la lecture des radiographies aux urgences (Figure 11). Parmi eux, 76,7% ont déclaré avoir déjà eu avoir accès à l'IA pour être aidé à lire des radiographies aux urgences.

Figure 10 : Formation et type de formation à la traumatologie



Note 1 : digramme de gauche 30 participants – diagramme de droite 13 participants
*Note 2 : * représente l'effectif en nombre*

Figure 11 : Expérience vis à vis des algorithmes d'IA d'aide à la lecture des radiographies aux urgences



Note 1 : 30 participants
*Note 2 : * représente l'effectif en nombre*

- **Performances des lecteurs**

1) Performance des lecteurs à l'échelle du cas

a) Sensibilité (Se)

Sur l'ensemble du jeu de données d'analyse, la Se de détection des lésions par les lecteurs tend à être supérieure avec l'aide de l'IA (68% vs 88%, et intervalles de confiance non chevauchants) (Tableau 3). Cela montre une amélioration significative de la capacité des lecteurs à détecter les cas positifs avec l'aide de l'IA.

On note également un gain de Se pour toutes les régions anatomiques mais le bénéfice de l'IA parait être particulièrement important pour détecter les lésions du poignet (66% vs 93%, IC non chevauchants) et de la cheville (73% vs 94%, IC non chevauchants) (Tableau 3).

Le bénéfice de l'aide de l'IA s'observe aussi chez les lecteurs du groupe « séniors » avec un gain de 23% et un bénéfice de 18% pour les lecteurs du groupe « internes » (Tableau 3).

`

b) Spécificité (Sp)

Notre jeu de données d'analyse incluait seulement 4 cas négatifs. Avec un petit effectif de cas négatifs, chaque cas négatif supplémentaire impacte fortement la spécificité et tout faux positif a tendance à faire fortement diminuer la Sp. Ainsi, avec un faible effectif global et par sous-groupes, et un effectif très réduit de cas négatifs, les données de Sp et leur interprétation sont à considérer avec beaucoup de précautions. On a aussi pu remarquer que les intervalles de confiance étaient larges autour de la Sp, ce qui rend la mesure imprécise (Tableau 4).

De plus, pour les sous-groupes de régions anatomiques ne présentant aucun cas négatif, le calcul de la spécificité n'a pas été possible. C'est pourquoi nous avons présenté le Tableau 4 sans l'interpréter.

Tableau 3 : Sensibilité des détections des lésions au global et par sous-groupe à l'échelle du cas

Groupe	Type	Se (sans IA)[1]	Se (avec IA)[1]	Delta Se[1]
Global (N=49)	**Global (N=49)**	0.683 [0.658 ; 0.708]	0.875 [0.857 ; 0.893]	0.192 [0.161 ; 0.223]
Régions anatomiques (N=49)	**Cheville (N=14)**	0.726 [0.683 ; 0.769]	0.936 [0.913 ; 0.959]	0.21 [0.161 ; 0.259]
	Coude (N=12)	0.603 [0.55 ; 0.656]	0.721 [0.673 ; 0.769]	0.118 [0.045 ; 0.191]
	Hanche (N=8)	0.758 [0.704 ; 0.812]	0.904 [0.867 ; 0.941]	0.146 [0.079 ; 0.213]
	Poignet (N=15)	0.656 [0.607 ; 0.705]	0.925 [0.898 ; 0.952]	0.269 [0.212 ; 0.326]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	0.685 [0.63 ; 0.74]	0.911 [0.877 ; 0.945]	0.226 [0.161 ; 0.291]
	Interne (N=49)	0.682 [0.654 ; 0.71]	0.866 [0.846 ; 0.886]	0.184 [0.149 ; 0.219]

[1]L'intervalle de confiance à 95% est entre crochet

Tableau 4 : Spécificité des détections des lésions au global et par sous-groupe à l'échelle du cas

Groupe	Type	Sp (sans IA)[1]	Sp (avec IA)[1]	Delta Sp[1]
Global (N=49)	**Global (N=49)**	0.55 [0.461 ; 0.639]	0.342 [0.257 ; 0.427]	-0.208 [-0.331 ; -0.085]
Régions anatomiques (N=49)	**Cheville (N=14)**	NaN [NaN ; NaN]	NaN [NaN ; NaN]	NaN [NaN ; NaN]
	Coude (N=12)	0.567 [0.39 ; 0.744]	0.167 [0.034 ; 0.3]	-0.4 [-0.621 ; -0.179]
	Hanche (N=8)	NaN [NaN ; NaN]	NaN [NaN ; NaN]	NaN [NaN ; NaN]
	Poignet (N=15)	0.544 [0.441 ; 0.647]	0.4 [0.299 ; 0.501]	-0.144 [-0.289 ; 0.001]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	0.625 [0.431 ; 0.819]	0.333 [0.144 ; 0.522]	-0.292 [-0.562 ; -0.022]
	Interne (N=49)	0.531 [0.431 ; 0.631]	0.344 [0.249 ; 0.439]	-0.187 [-0.324 ; -0.05]

[1]NaN : Non applicable

[1]L'intervalle de confiance à 95% est entre crochet

c) Valeur prédictive positive (VPP)

Sur l'ensemble du jeu de données d'analyse, la VPP sans IA était de 0.945 [0.931 ; 0.959], la VPP avec IA de 0.937 [0.924 ; 0.95], donc une différence de -0.008 [-0.028 ; 0.012]. Cependant, on a remarqué que les intervalles de confiance des VPP avec et sans IA se chevauchaient, ce qui signifie qu'il y a une incertitude quant à la différence réelle entre ces 2 valeurs de VPP. On peut difficilement affirmer que l'IA a un réel impact sur la VPP étant donné que ces 2 intervalles partagent des valeurs possibles. Ce même constat est fait dans tous les sous-groupes de régions anatomiques et dans les sous-groupes d'expérience (Tableau 5).

La VPP pour les cas de hanche et de cheville était parfaite car il n'y avait aucun cas négatif dans ces 2 sous-groupes. On ne peut donc tirer aucune conclusion sur ces 2 groupes.

Dans le sous-groupe poignet, bien qu'il existe une incertitude quant à la différence de mesure avec et sans IA, ce groupe paraissait être le seul où l'IA a eu un impact positif avec un gain de 0,08% (Tableau 5).

Tableau 5 : Valeur prédictive positive (VPP) des détections des lésions au global et par sous-groupes à l'échelle du cas

Groupe	Type	VPP (sans IA)[1]	VPP (avec IA)[1]	Delta VPP[1]
Global (N=49)	**Global (N=49)**	0.945 [0.931 ; 0.959]	0.937 [0.924 ; 0.95]	-0.008 [-0.028 ; 0.012]
Régions anatomiques (N=49)	**Cheville (N=14)**	1 [1 ; 1]	1 [1 ; 1]	0 [0 ; 0]
	Coude (N=12)	0.939 [0.907 ; 0.971]	0.905 [0.87 ; 0.94]	-0.034 [-0.081 ; 0.013]
	Hanche (N=8)	1 [1 ; 1]	1 [1 ; 1]	0 [0 ; 0]
	Poignet (N=15)	0.852 [0.81 ; 0.894]	0.86 [0.825 ; 0.895]	0.008 [-0.047 ; 0.063]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	0.954 [0.925 ; 0.983]	0.939 [0.91 ; 0.968]	-0.015 [-0.056 ; 0.026]
	Interne (N=49)	0.942 [0.926 ; 0.958]	0.937 [0.922 ; 0.952]	-0.005 [-0.027 ; 0.017]

[1]L'intervalle de confiance à 95% est entre crochet

d) Valeur prédictive négative (VPN)

De la même manière que pour la spécificité, il nous a été impossible d'interpréter correctement la mesure de la VPN. Le faible effectif de cas négatif, 4 seulement, ne nous permet pas de conclure avec assurance de l'impact de l'IA sur la VPN. Les résultats de la mesure de la VPN sont rapportés dans le Tableau 6 mais ils ne sont pas représentatifs ou peu fiable en raison de notre faible échantillon de cas négatifs.

Tableau 6 : Valeur prédictive négative (VPN) des détections des lésions au global et par sous-groupes à l'échelle du cas

Groupe	Type	VPN (sans IA)[1]	VPN (avec IA)[1]	Delta VPN[1]
Global (N=49)	**Global (N=49)**	0.134 [0.104 ; 0.164]	0.195 [0.141 ; 0.249]	0.061 [0 ; 0.122]
Régions anatomiques (N=49)	**Cheville (N=14)**	0 [0 ; 0]	0 [0 ; 0]	0 [0 ; 0]
	Coude (N=12)	0.115 [0.064 ; 0.166]	0.052 [0.008 ; 0.096]	-0.063 [-0.132 ; 0.006]
	Hanche (N=8)	0 [0 ; 0]	0 [0 ; 0]	0 [0 ; 0]
	Poignet (N=15)	0.15 [0.08 ; 0.22]	0.25 [0.1 ; 0.4]	0.1 [-0.065 ; 0.265]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	0.129 [0.096 ; 0.162]	0.185 [0.128 ; 0.242]	0.056 [-0.011 ; 0.123]
	Interne (N=49)	0.13 [0.097 ; 0.163]	0.187 [0.13 ; 0.244]	0.057 [-0.008 ; 0.122]

[1]L'intervalle de confiance à 95% est entre crochet

e) Temps de lecture

Nous avons aussi analysé le temps de lecture de chaque lecteur avec et sans l'aide de l'IA. De manière globale nous avons remarqué une différence entre les temps moyen et médian de lecture mais aussi par sous-groupe de région anatomique et d'expérience. Ceci pourrait s'expliquer par des valeurs extrêmes supérieures (temps de lecture plus longs) qui tirent la moyenne vers le haut. Globalement et dans tous les sous-groupes, nous avons remarqué une amélioration du temps médian et moyen de lecture, avec un effet un plus important sur le temps moyen, ce qui pourrait laisser croire que l'IA a un peu moins d'effet sur les valeurs extrêmes. Nous avons rapporté les résultats de ces 2 mesures dans le Tableau 7 et le Tableau 8. Nous avons aussi présenté les résultats sous forme de box-plot[i] (Figure 12, Figure 13, Figure 14, Figure 18).

Globalement, la médiane sans IA était de 88,95 secondes [83,72 ; 94,18] et la médiane avec IA de 48,08 secondes [43,53 ; 52,63], soit une différence de 40,87 secondes. On peut donc dire que l'IA parait avoir un impact sur la diminution du temps de lecture (Tableau 8). La Figure 12 montre qu'il existait une réduction du temps médian de lecture lorsqu'on utilisait l'IA. En effet, la médiane du temps de lecture avec Milvue Suite était inférieure à celle sans Milvue Suite, soulignant l'efficacité accrue apportée par l'IA. On a observé également que la boîte verte (avec Milvue Suite) était plus compacte que la boîte rouge (sans Milvue Suite), ce qui indique

[i] Aussi appelé « boite à moustache ». La définition est donnée en annexe 3.

une réduction de la variabilité des temps de lecture. Cela suggère que l'IA rend les performances des lecteurs plus homogènes. Après utilisation de Milvue Suite, on retrouvait encore des valeurs aberrantes mais elles étaient moins fréquentes et moins extrêmes. Cela pourrait montrer que l'IA serait utile pour réduire le nombre de cas complexes à lire (Figure 12). Par zones anatomiques, on a retrouvé des résultats similaires (Tableau 7, Tableau 8, Figure 14), avec une réduction du temps de lecture avec l'utilisation de l'IA dans chaque sous-groupe. Concernant les sous-groupes d'expérience, nous avons observé une diminution du temps médian de lecture (35.56 secondes [34.08 ; 37.04] pour les séniors et 40.46 [39.63 ; 41.29] pour les internes). Nous avons également observé que les séniors avaient des valeurs de temps de lecture plus homogènes (différence modérée entre la moyenne et la médiane) : moyenne sans IA à 102,31 secondes et une médiane sans IA à 70,26 secondes. Cela suggère qu'ils étaient généralement plus constants dans leurs temps d'analyse. L'IA aide à réduire de manière importante le temps de lecture, surtout pour les cas longs à analyser, comme le montre la réduction plus importante de la moyenne par rapport à la médiane (Tableau 7, Tableau 8 et Figure 13). Les internes en revanche, semblaient avoir plus de variabilité dans leurs temps de lecture, comme en témoigne l'écart plus important entre la moyenne et la médiane (moyenne sans IA : 126,83 secondes ; médiane sans IA : 94,81 secondes). Cela pourrait refléter leur moindre expérience, où certains cas prennent beaucoup plus de temps à analyser. L'IA montrait un impact positif sur leur temps de lecture médian et moyen (moyenne avec IA : 75,46 secondes ; médiane avec IA : 54,35 secondes ; delta moyenne : 51,37 secondes ; delta médiane : 40,46 secondes), bien qu'il semblait persister quelques cas difficiles à lire (Tableau 7, Tableau 8, Figure 13).

Tableau 7 : Temps moyen de lecture au global et par sous-groupe

Groupe	Type	Temps moyen en secondes sans IA[1]	Temps moyen en secondes avec IA[1]	Delta temps moyen[1]
Global (N=49)	**Global (N=49)**	122.3 [117.07 ; 127.53]	69.37 [64.82 ; 73.92]	52.93 [52.2 ; 53.66]
Régions anatomiques (N=49)	**Cheville (N=14)**	124.53 [115.31 ; 133.75]	78.68 [71.09 ; 86.27]	45.85 [44.55 ; 47.15]
	Coude (N=12)	108.9 [97.59 ; 120.21]	69.55 [59 ; 80.1]	39.35 [37.75 ; 40.95]
	Hanche (N=8)	132.25 [119.93 ; 144.57]	68.41 [56.7 ; 80.12]	63.84 [61.98 ; 65.7]
	Poignet (N=15)	123.56 [113.75 ; 133.37]	58.8 [50.57 ; 67.03]	64.76 [63.48 ; 66.04]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	102.31 [90.14 ; 114.48]	45.9 [35.88 ; 55.92]	56.41 [54.93 ; 57.89]
	Interne (N=49)	126.83 [121.04 ; 132.62]	75.46 [70.36 ; 80.56]	51.37 [50.54 ; 52.2]

[1]L'intervalle de confiance à 95% est entre crochet

Tableau 8 : Temps médian de lecture au global et par sous-groupe

Groupe	Type	Temps médian en secondes sans IA	Temps médian en secondes avec IA	Delta temps médian
Global (N=49)	**Global (N=49)**	88.95 [83.72 ; 94.18]	48.08 [43.53 ; 52.63]	40.87 [40.14 ; 41.6]
Régions anatomiques (N=49)	**Cheville (N=14)**	89.3 [80.08 ; 98.52]	57.22 [49.63 ; 64.81]	32.08 [30.78 ; 33.38]
	Coude (N=12)	79.82 [68.51 ; 91.13]	44.1 [33.55 ; 54.65]	35.72 [34.12 ; 37.32]
	Hanche (N=8)	100.71 [88.39 ; 113.03]	48.08 [36.37 ; 59.79]	52.63 [50.77 ; 54.49]
	Poignet (N=15)	93.53 [83.72 ; 103.34]	39.66 [31.43 ; 47.89]	53.87 [52.59 ; 55.15]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	70.26 [58.09 ; 82.43]	34.7 [24.68 ; 44.72]	35.56 [34.08 ; 37.04]
	Interne (N=49)	94.81 [89.02 ; 100.6]	54.35 [49.25 ; 59.45]	40.46 [39.63 ; 41.29]

[1]L'intervalle de confiance à 95% est entre crochet

Figure 12 : Temps de lecture avec sans utilisation de Milvue Suite (au global)

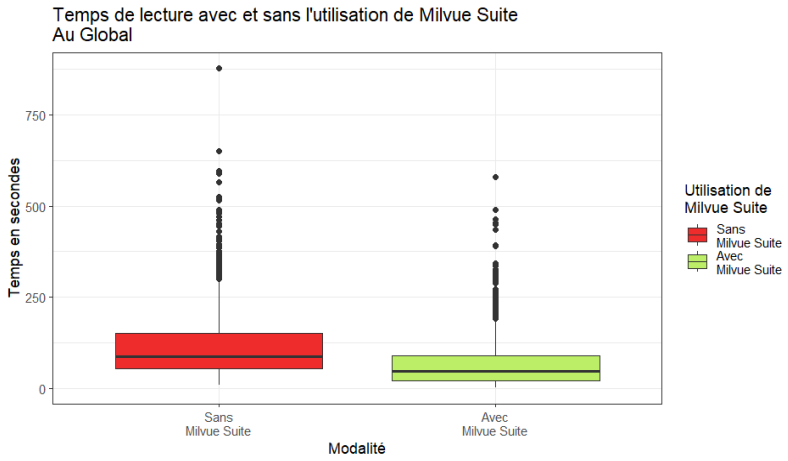


Figure 13 : Temps de lecture avec et sans utilisation de Milvue Suite (par expérience des lecteurs)

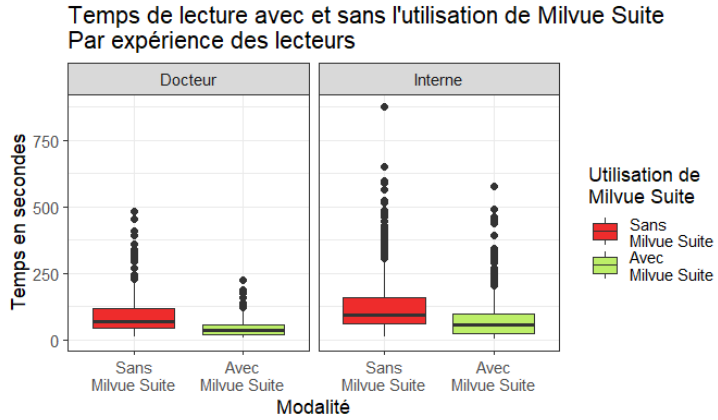
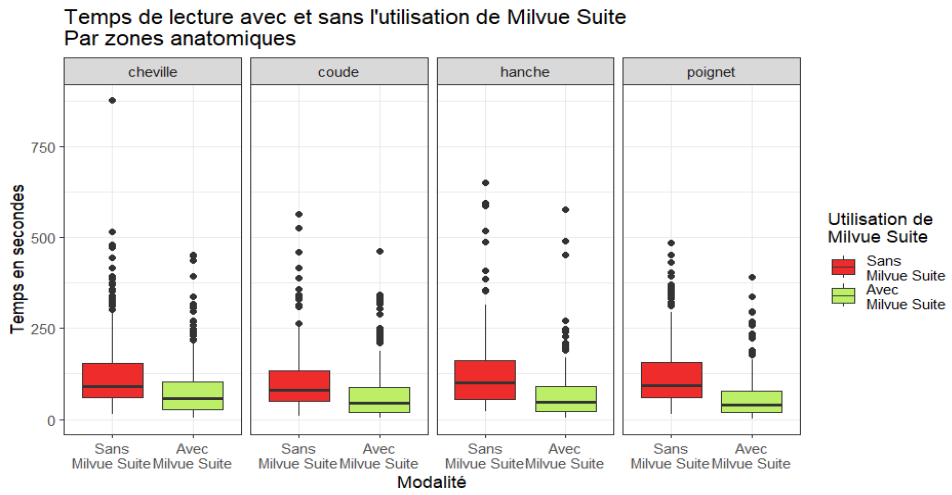


Figure 14 : Temps de lecture avec sans utilisation de Milvue Suite (par zone anatomique)



2) Performance des lecteurs à l'échelle de la lésion

a) Sensibilité (Se)

Globalement, l'IA a amélioré considérablement la sensibilité des lecteurs à l'échelle des lésions, avec une augmentation de 28 %. Cela signifie que les lecteurs détectaient un plus grand nombre de lésions individuelles lorsqu'ils étaient aidés de l'IA, réduisant le risque de passer à côté de lésion importantes. Cette amélioration était encore plus marquée qu'à l'échelle du cas (19%), ce qui montre que l'IA était particulièrement performante pour identifier les lésions individuelles.

Dans tous les sous-groupes de régions anatomiques, la sensibilité a été améliorée avec l'aide de Milvue Suite (Tableau 9). On a remarqué une franche amélioration sur le poignet (0.489 à 0.912 ; +0.423) et pour le coude (0.451 à 0.716 ; +0.265). On pourrait dire que l'algorithme est particulièrement utile dans des régions anatomiques plus complexes où la détection des lésions est difficile.

De la même manière, dans les sous-groupes d'expérience, la sensibilité a été améliorée avec l'aide de l'IA (Tableau 9). La sensibilité du groupe « sénior » a augmentée de 0.567 à 0.868 (+0.301) et celle du groupe « interne » de 0.576 à 0.854 (+0.278). Milvue suite parait être aussi utile aux lecteurs expérimentés que moins expérimentés dans la détection des lésions individuelles.

Tableau 9 : Sensibilité (Se) des détections de lésions au global et par sous-groupes à l'échelle de la lésion

Groupe	Type	Se (sans IA)[1]	Se (avec IA)[1]	Delta Se[1]
Global (N=49)	**Global (N=49)**	0.575 [0.561 ; 0.589]	0.857 [0.846 ; 0.868]	0.282 [0.264 ; 0.3]
Régions anatomiques (N=49)	**Cheville (N=14)**	0.695 [0.673 ; 0.717]	0.921 [0.908 ; 0.934]	0.226 [0.201 ; 0.251]
	Coude (N=12)	0.451 [0.415 ; 0.487]	0.716 [0.682 ; 0.75]	0.265 [0.216 ; 0.314]
	Hanche (N=8)	0.605 [0.566 ; 0.644]	0.73 [0.694 ; 0.766]	0.125 [0.072 ; 0.178]
	Poignet (N=15)	0.489 [0.465 ; 0.513]	0.912 [0.896 ; 0.928]	0.423 [0.394 ; 0.452]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	0.567 [0.535 ; 0.599]	0.868 [0.845 ; 0.891]	0.301 [0.262 ; 0.34]
	Interne (N=49)	0.576 [0.56 ; 0.592]	0.854 [0.842 ; 0.866]	0.278 [0.258 ; 0.298]

[1]*L'intervalle de confiance à 95% est entre crochet*

b) Valeur prédictive positive (VPP)

Bien que l'IA améliorait la détection des lésions (amélioration de la Se à l'échelle des lésions, elle entrainait une diminution notable de la VPP de manière générale (0.803 sans IA vs 0.704 avec IA soit différence de -0.099) (Tableau 10). Autrement dit, l'IA générait plus de faux positifs à l'échelle des lésions et donc un risque de sur-diagnostique.

Au niveau des sous-groupes anatomiques, il est difficile de statuer sur la VPP de la hanche et du coude dans la mesure où les IC avant et après IA se chevauchaient. Cependant pour les autres régions, les résultats semblaient plus francs. Dans le sous-groupe cheville, la VPP était réduite de 0.913 à 0.709 (-0.204) et la VPP était réduite de 0.791 à 0.74 (-0.051) pour le sous-groupe poignet (Tableau 10). On pourrait dire que l'IA n'améliorait pas beaucoup la proportion de vrais positifs parmi les résultats détectés, mais elle ne dégradait pas non plus de manière importante la performance globale.

Les résultats étaient similaires dans les 2 sous-groupes d'expérience, avec un impact plus prononcé chez les « séniors ». Chez ces derniers, la VPP était réduite de 0.872 à 0.673 (-0.199) et les internes la VPP a été réduite de 0.787 à 0.713 (-0.074) (Tableau 10). Cela pourrait indiquer que les seniors, malgré leur expérience, pourraient être influencés par l'IA pour sur-diagnostiquer certaines lésions, tandis que les internes, qui bénéficient d'une plus grande assistance de l'IA, voient une réduction moindre de la VPP.

Tableau 10 : Valeur prédictive positive (VPP) des détections de lésions au global et par sous-groupes à l'échelle de la lésion

Groupe	Type	VPP (sans IA)[1]	VPP (avec IA)[1]	Delta VPP[1]
Global (N=49)	**Global (N=49)**	0.803 [0.79 ; 0.816]	0.704 [0.691 ; 0.717]	-0.099 [-0.117 ; -0.081]
Régions anatomiques (N=49)	**Cheville (N=14)**	0.913 [0.898 ; 0.928]	0.709 [0.69 ; 0.728]	-0.204 [-0.229 ; -0.179]
	Coude (N=12)	0.636 [0.594 ; 0.678]	0.59 [0.557 ; 0.623]	-0.046 [-0.099 ; 0.007]
	Hanche (N=8)	0.708 [0.669 ; 0.747]	0.755 [0.72 ; 0.79]	0.047 [-0.006 ; 0.1]
	Poignet (N=15)	0.791 [0.766 ; 0.816]	0.74 [0.718 ; 0.762]	-0.051 [-0.084 ; -0.018]
Expérience des lecteurs (N=49)	**Sénior (N=49)**	0.872 [0.845 ; 0.899]	0.673 [0.645 ; 0.701]	-0.199 [-0.238 ; -0.16]
	Interne (N=49)	0.787 [0.772 ; 0.802]	0.713 [0.699 ; 0.727]	-0.074 [-0.096 ; -0.052]

[1]L'intervalle de confiance à 95% est entre crochet

Il est intéressant de comparer les résultats à l'échelle du cas et de la lésion (Tableau 9 et Tableau 10). La sensibilité à l'échelle du cas (0.683) était généralement plus élevée sans IA que la sensibilité à l'échelle de la lésion sans IA (0.575). Cela s'explique par le fait qu'il est plus facile pour les lecteurs de détecter au moins une lésion dans un cas global que de repérer chaque lésion individuelle. Mais avec l'IA, la sensibilité à l'échelle de la lésion atteignait pratiquement le même niveau que la sensibilité à l'échelle du cas (près de 88% et 86% respectivement) avec un impact plus fort à l'échelle de la lésion (+ 0,282). L'IA aiderait considérablement les lecteurs à identifier plus de lésions individuelles dans chaque image, ce qui est primordial pour des diagnostics plus précis et complets. Concernant la VPP, à l'échelle du cas il a été difficile de l'interpréter, la mesure étant incertaine. Cependant la mesure de la VPP avant et après utilisation de l'IA paraissait tendre vers une légère diminution. A l'échelle de la lésion, nous pouvions interpréter cette mesure avec plus de certitude, et nous avons remarqué la aussi une diminution peu importante (- 0,099). L'aide de l'IA parait générer plus de faux positifs à l'échelle de la lésion.

- **Performances de l'algorithme**

1) *Performance globale*

L'interprétation des courbes ROC de l'algorithme Milvue Suite dans la détection des fractures, des luxations et des pathologies sont à prendre avec précaution (Figure 15, Figure 16 et Figure 17). Rappelons que le jeu de données d'analyse était biaisé (45 cas positifs et 4 cas négatifs), ce qui aurait pu surestimer les performances diagnostiques de l'algorithmes.
La courbe ROC de détection des fractures montrait une AUC de 0,92, ce qui reste indicatif d'une bonne performance de l'algorithme (Figure 15). Le fait que l'intervalle de confiance était relativement étroit (0,882-0,959) renforçait la fiabilité de ce résultat. Cependant, l'existence de peu de cas négatifs aurait pu exagérer cette performance. L'algorithme pouvait être fortement influencé par la capacité à détecter correctement les cas positifs, mais pouvait ne pas être suffisamment testé sur les cas négatifs. Ceci signifie que l'algorithme parait être performant pour détecter des fractures, mais son potentiel pour éviter les faux positifs est moins clair, et pourrait être surestimé par cette AUC.
La courbe ROC de détection des luxations montrait une AUC quasi parfaite de 0,99 avec un IC étroit (0,98 ;1) (Figure 16). De la même manière que pour la courbe ROC précédente, celle-ci est à interpréter prudemment. L'AUC pouvait également être biaisée par la surreprésentation des cas positifs. La performance presque parfaite de l'algorithme dans la détection des luxations aurait pu être en réalité moins robuste si l'algorithme avait été testé avec un nombre plus important de cas négatifs.

Figure 15 : Courbe ROC de l'algorithme Milvue Suite dans la détection des fractures

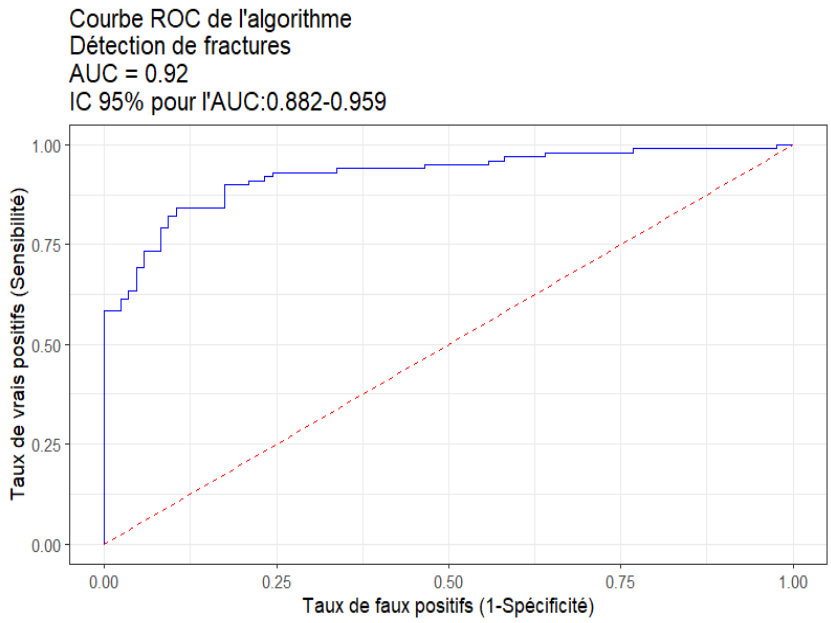
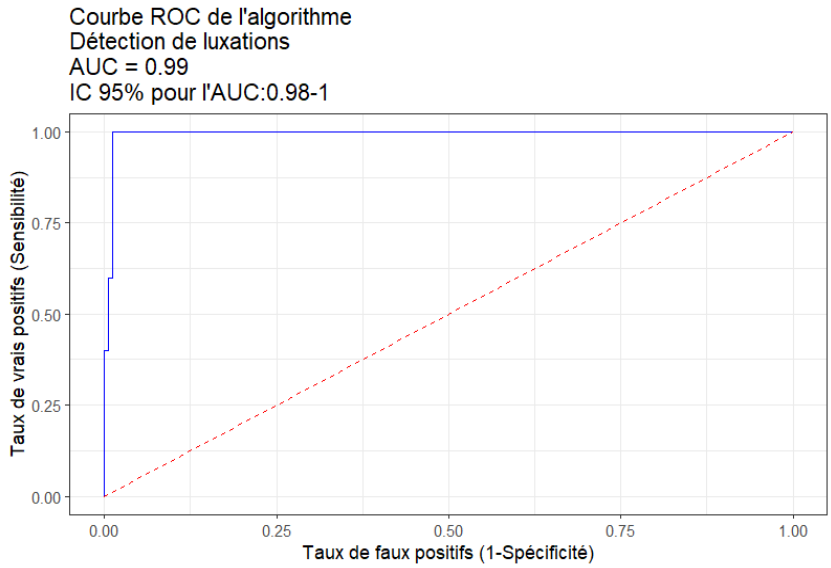
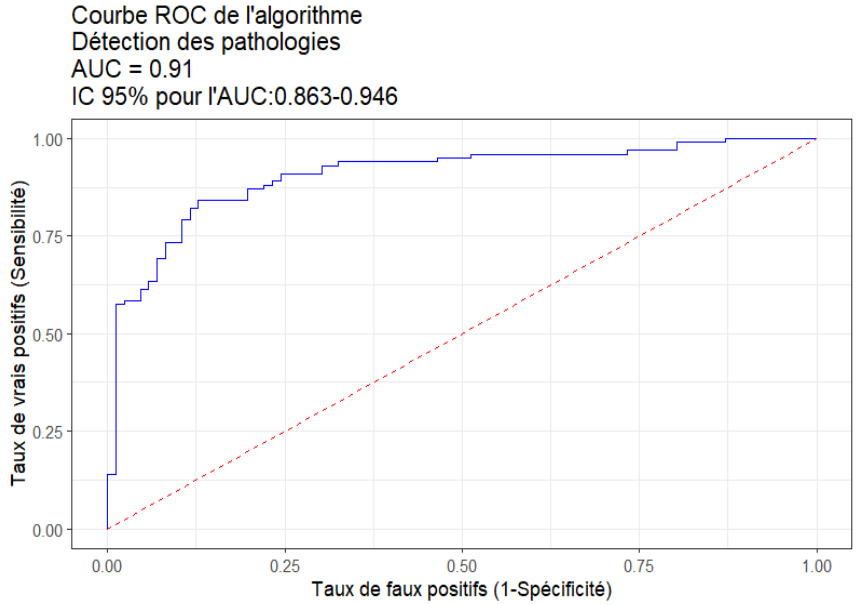


Figure 16 : Courbe ROC de l'algorithme Milvue Suite dans la détection des luxations



La courbe ROC de détection des pathologies (luxation + fracture) montrait une AUC de 0,91 avec un intervalle de confiance assez étroit (0,863 ;0,946) (Figure 17). L'algorithme Milvue Suite montrait une bonne performance globale pour détecter les pathologies, bien que le jeu de données d'analyse fût déséquilibré.

Figure 17 : Courbe ROC de l'algorithme Milvue Suite dans la détection des pathologies



2) *Performance à l'échelle du cas*

a) Sensibilité (Se)

Globalement, la Se était élevée (0,911, IC [0.828 ; 0.994]), ce qui montrait que l'algorithme était capable de détecter la plupart des lésions (Tableau 11).
Concernant l'analyse en sous-groupe, la Se des groupes « cheville » et « poignet » était de 1.00. L'algorithme a donc détecté toutes les lésions dans ces sous-groupes.
Les Se des groupes « coude » et « hanche » sont à interpréter avec précaution car les IC étaient larges, ce qui rendait la mesure peu précise (0,727, IC [0.464 ; 0.99] et 0.875 [0.646 ; 1] respectivement). L'algorithme paraissait détecter une majorité de lésion, mais les résultats semblaient être très variable (largeur des IC) (Tableau 11).

Tableau 11 : Sensibilité (Se) des détections de pathologies (fractures et luxations) au global et par sous-groupes à l'échelle du cas

Groupe	Type[1]	Se (IA)[1]
Global (N=49)	**Global (N=49)**	0.911 [0.828 ; 0.994]
Régions anatomiques (N=49)	**Cheville (N=14)**	1 [1 ; 1]
	Coude (N=12)	0.727 [0.464 ; 0.99]
	Hanche (N=8)	0.875 [0.646 ; 1]
	Poignet (N=15)	1 [1 ; 1]

[1]L'intervalle de confiance à 95% est entre crochet

b) Spécificité (Sp)

A l'image de l'analyse des performances des lecteurs, il n'a pas été possible d'interpréter correctement les résultats de la spécificité de l'algorithme à détecter des lésions. Notre jeu de données d'analyse ne contenait pas assez de cas négatifs pour pouvoir conclure sur cette mesure. Nous avons donné les résultats dans le Tableau 12 mais nous n'avons pas interpréter les résultats.

Tableau 12 : Spécificité (Sp) des détections de pathologies (fractures et luxations) au global et par sous-groupes l'échelle du cas

Groupe	Type[1]	Sp (IA)[1]
Global (N=49)	**Global (N=49)**	0.25 [0 ; 0.674]
Régions anatomiques (N=49)	**Cheville (N=14)**	NaN [NaN ; NaN]
	Coude (N=12)	0 [0 ; 0]
	Hanche (N=8)	NaN [NaN ; NaN]
	Poignet (N=15)	0.333 [0 ; 0.866]

[1]NaN : Non applicable
[1]L'intervalle de confiance à 95% est entre crochet

c) Valeur prédictive positive (VPP)

Globalement, la VPP des détections des lésions par l'algorithme était naturellement bonne 0.932 [0.858 ; 1] par la prédominance de cas positifs dans notre jeu de données d'analyse (Tableau 13). Ce résultat suggérait que la majorité des diagnostics positifs fait par le logiciel Milvue Suite étaient corrects.
Par sous-groupes de régions anatomiques, les IC étaient plus larges (Tableau 13), ce qui rend la mesure de la VPP peu précise et son interprétation difficile à faire. Le sous-groupe « cheville » et « hanche » avait naturellement une VPP égale à 1, étant donné qu'il n'y avait aucun cas négatif. On peut dire que par sous-groupes de régions anatomiques, l'algorithme tendait à diagnostiquer de vrais cas positifs mais que cette mesure semblait être variable.

d) Valeur prédictive négative (VPN)

A propos de la VPN de l'algorithme à détecter les lésions à l'échelle du cas, il est peu pertinent d'interpréter les résultats. En effet, au global, l'IC est très large 0.2 [0 ; 0.551] donc la mesure peu précise, et pour le sous-groupe « cheville » cette mesure n'a pas pu être calculé (Tableau 14). A propos du poignet, l'algorithme semblait avoir correctement diagnostiqué les cas où il y avait effectivement une lésion. A propos du coude, l'algorithme semblait ne pas avoir correctement diagnostiqué les cas réellement négatifs. A propos de la hanche, l'algorithme semblait avoir réussi à exclure la totalité des cas positifs en diagnostiquant des faux négatifs.

Tableau 13 : Valeur prédictive positive (VPP) des détections de pathologies (fractures et luxations) au global et par sous-groupes à l'échelle du cas

Groupe	Type[1]	VPP (IA)[1]
Global (N=49)	**Global (N=49)**	0.932 [0.858 ; 1]
Régions anatomiques (N=49)	**Cheville (N=14)**	1 [1 ; 1]
	Coude (N=12)	0.889 [0.684 ; 1]
	Hanche (N=8)	1 [1 ; 1]
	Poignet (N=15)	0.857 [0.674 ; 1]

[1]*L'intervalle de confiance à 95% est entre crochet*

Tableau 14 : Valeur prédictive négative (VPN) des détections de pathologies (fractures et luxations) au global et par sous-groupes à l'échelle du cas

Groupe	Type[1]	VPN (IA)[1]
Global (N=49)	**Global (N=49)**	0.2 [0 ; 0.551]
Régions anatomiques (N=49)	**Cheville (N=14)**	NaN [NaN ; NaN]
	Coude (N=12)	0 [0 ; 0]
	Hanche (N=8)	0 [0 ; 0]
	Poignet (N=15)	1 [1 ; 1]

[1]*NaN : Non applicable*

[1]*L'intervalle de confiance à 95% est entre crochet*

3) Performance à l'échelle de la lésion

a) Sensibilité (Se)

La valeur de Se état globalement bonne (0.857 [0.799 ; 0.915]), et les valeurs de Se dans la plupart des sous-groupes de régions anatomiques aussi (Tableau 15).
On a remarqué que les IC des sous-groupes « coude » et « hanche » étaient très larges, les valeurs de Se de ces groupes étaient donc peu précises. L'algorithme apparaissait moins fiable dans ces régions pour identifier les lésions (Tableau 15).

Tableau 15 : Sensibilité (Se) des détections de pathologies (fractures et luxations) au global et par sous-groupes à l'échelle de la lésion

Groupe	Type[1]	Se (IA)[1]
Global (N=49)	**Global (N=49)**	0.857 [0.799 ; 0.915]
Régions anatomiques (N=49)	**Cheville (N=14)**	0.909 [0.833 ; 0.985]
	Coude (N=12)	0.739 [0.56 ; 0.918]
	Hanche (N=8)	0.7 [0.499 ; 0.901]
	Poignet (N=15)	0.929 [0.851 ; 1]

[1]*L'intervalle de confiance à 95% est entre crochet*

b) Valeur prédictive positive (VPP)

La VPP de l'algorithme dans la détection des lésions à l'échelle de la lésion semblait globalement bonne et fiable, avec un IC étroit (0.727 [0.659 ; 0.795]) (Tableau 16). Dans les ¾ des cas, l'algorithme à correctement diagnostiqué les cas positifs.

Par sous-groupes de régions anatomiques, les valeurs de VPP semblaient aussi être plus contrastée. Les groupes « coude » et hanche » avaient des IC très large autour des mesures de VPP, ce qui les rendaient peu précise bien qu'elles apparaissaient élevées, notamment pour la hanche (Tableau 16). On peut dire que l'algorithme semblait diagnostiqué pour la grande majorité des sous-groupes correctement les cas réellement positifs, mais dans certains sous-groupes, comme la hanche et le coude, il paraissait moins précis.

Tableau 16 : Valeur prédictive positive (VPP) des détections de pathologies (fractures et luxations) au global et par sous-groupes à l'échelle de la lésion

Groupe	Type[1]	VPP (IA)[1]
Global (N=49)	**Global (N=49)**	0.727 [0.659 ; 0.795]
Régions anatomiques (N=49)	**Cheville (N=14)**	0.725 [0.62 ; 0.83]
	Coude (N=12)	0.63 [0.448 ; 0.812]
	Hanche (N=8)	0.824 [0.643 ; 1]
	Poignet (N=15)	0.75 [0.632 ; 0.868]

[1]*L'intervalle de confiance à 95% est entre crochet*

DISCUSSION

- **Impact de l'IA sur les performances diagnostiques**

Dans cette étude, nous avons essayé de montrer quel était l'impact de la pré analyse des radiographies traumatologiques par l'algorithme d'intelligence artificielle Milvue Suite, aux urgences de Maubeuge. Nous n'avons trouvé aucune étude comparant les performances de lecture à la fois d'urgentistes séniors et internes, avec et sans IA. Aucune étude s'intéressant aux performances de lecture avec l'aide l'IA chez les urgentistes utilisaient un ground-truth scanographique.

Les résultats ont montré une amélioration de la sensibilité grâce à l'IA, dans la détection des lésions de type fractures/luxations à l'échelle du cas et de la lésion. Ces résultats sont en accord avec l'étude de Duron et al. en 2021 (27), qui a également observé une augmentation de la sensibilité (+8,7%) et de la spécificité (+4,1%) chez les radiologues et les urgentistes utilisant une aide à l'interprétation des fractures appendiculaires.

L'AUC de notre algorithme calculée à 0,91, était en accord avec les résultats de plusieurs études, notamment par l'étude de Loïc Duron et son équipe (27) qui avaient aussi mesuré une AUC à 0,91. Durant leurs travaux, un nouveau logiciel avait été développé avec une AUC allant même jusqu'à 94%. Concernant les performances de l'IA dans notre étude, malgré notre faible effectif, il n'y a pas eu de risque d'overfitting ou surajustement. En effet l'IA utilisée dans l'étude, a été entrainé sur la base de données de Milvue et non sur notre faible échantillon. Nous avons montré que la Se de l'algorithme Milvue Suite dans la détection des lésions de type fracture ou luxation pouvait dépasser les performances de certains cliniciens et notamment des plus jeunes (91% vs 87%). De manière similaire, Jones et al. en 2020 (28) ont montré que leur système d'apprentissage profond pouvait atteindre une sensibilité de 95,2% pour la détection des fractures musculosquelettiques, surpassant là aussi les performances de certains praticiens expérimentés. La Se dans la détection des lésions avec l'aide de l'IA a été à la fois amélioré chez les internes et les séniors dans notre études (+18% et +22%). L'étude du français T. Jacques (25), qui comparait les performances de lectures de radiologues séniors et juniors, en se basant elle aussi sur un ground-truth scanographique, a montré une amélioration de la Se dans son groupe junior et sénior. L'impact sur la Se était cependant moins important que dans notre étude (+6,1% et +3,4%), ce qui pourrait être s'expliquer par le fait que les lecteurs dans cette étude étaient experts de la discipline. Les récents travaux de T. Fu et son équipe (29), ont montré une augmentation significative de la Se avec l'aide de l'IA, lors de la détection de fractures des extrémités sur des radiographies (87% vs 96%). Ces résultats confirment que l'IA peut être un outil intéressant pour améliorer les performances diagnostique dans la lecture des

radiographies traumatologiques, et ce, à la fois pour les séniors et les internes non experts de la radiographie (à l'image des urgentistes).

Nos résultats ont aussi permis de montrer une réduction de temps de lecture, à la fois pour les internes et pour les séniors, ce qui est cohérent avec les conclusions de Duron et al (27), qui ont observé une réduction du temps de lecture de 15% en moyenne. Nos résultats font écho aux résultats de l'étude de T. Fu (29). Cette étude de 2024, comparait les performances de lectures de radiologues et urgentistes devant la détection de fractures des extrémités, avec l'aide de l'IA et sans l'aide de l'IA. L'étude a montré une réduction de temps de lecture de 27%. Ce gain sur de temps est particulièrement pertinent dans les services d'urgence, où la rapidité est essentielle pour la prise en charge des patients. Shin et al (30) ont aussi montré une réduction de temps de lecture grâce à l'aide de l'IA, sur des radiographies thoraciques cette fois. Laur et Wang dans leur revue de 2022 (31) ont eux aussi montré que l'IA permettait, en plus d'améliorer la précision diagnostique des lecteurs non experts en radiologie, une réduction du temps de lecture et ce sur différentes techniques d'imagerie dont la radiographie. Une étude de Waymel et al en 2019 (32) a permis de montré qu'un des souhaits des praticiens en utilisant l'IA en radiologie, était de diminuer le temps de lecture et donc d'augmenter le temps passé auprès des patients. Ces résultats suggèrent que le temps gagné à interpréter des radiographies grâce l'IA permettrait de passer plus de temps auprès du patient en augmentant le « temps clinique ».

L'ensemble de ces résultats sur la Se et le temps de lecture suggèrent que l'IA pourrait améliorer l'efficacité du flux de travail dans les urgences, permettant de diagnostiquer un plus grand nombre de lésions en moins de temps.

Cependant, dans notre étude, la VPP à l'échelle de la lésion diminuait légèrement dans tous les groupes d'analyse et notamment dans le sous-groupe « sénior ». Pour la VPP à l'échelle du cas, il nous a été difficile de statuer sur l'impact de l'IA, devant une incertitude quant à la mesure. Mais les résultats semblaient aller vers une diminution de la valeur prédictive positive. Il parait donc important d'utiliser l'IA avec prudence afin de limiter le risque de sur-diagnostique. En comparant la sensibilité à l'échelle du cas et de la lésion, nous sommes rendus compte que Milvue Suite avait un impact considérable en pratique. À l'échelle du cas, l'IA semble être un outil fiable pour améliorer la détection des cas positifs sans introduire trop de faux positifs, ce qui est utile pour des diagnostics rapides dans un environnement comme les urgences, où le flux de patient est important. À l'échelle de la lésion, l'IA pourrait être particulièrement utile pour des contextes où il est essentiel de ne manquer aucune lésion (notamment pour décider d'un traitement), mais cela exige également un suivi clinique rigoureux pour vérifier les détections, en raison du risque accru de faux positifs.

- **Les limites de l'étude**

Notre étude avait de nombreuses limites. Premièrement, elle était rétrospective, comme la plupart des études sur le même thème (33,34), et monocentrique. Les résultats ne sont donc pas représentatifs de l'ensemble des CH de France mais uniquement du CH de Maubeuge. Deuxièmement, aucun test statistique n'a été fait lors de l'analyse des résultats. Il n'est donc pas possible de conclure à une significativité statistique des résultats, nous pouvons seulement évoquer des tendances. L'une des principales limitations de cette étude était la taille restreinte de l'échantillon, avec seulement 49 clichés analysés, et un nombre très limité de cas négatifs (4 cas). Cette taille d'échantillon réduit la validité et la fiabilité de nos résultats et augmente la sensibilité des résultats à des variations aléatoires. Par conséquent, bien que nos résultats suggèrent une amélioration de la détection des anomalies grâce à l'IA, ces conclusions doivent être interprétées avec prudence et confirmées par des études futures avec un plus grand échantillon. En outre, le faible nombre de cas négatifs a entrainé un biais de sélection et de mesure ; et ne nous a pas permis d'interpréter la spécificité et la valeur prédictive négative. Nos résultats semblaient être discordants avec plusieurs études ayant retrouvées une augmentation de la Sp avec l'aide de l'IA ou bien une non-diminution de la Sp. C'est par exemple le cas de l'étude française de Lisa Canoni-Meynet (35), certes sur des lecteurs radiologues, mais qui montre à la fois une augmentation de la Se, de la Sp, de la VPP et de la VPN avec l'aide de l'IA (+20%, +0,6%, +2,9%, +10% respectivement). L'étude de l'équipe de Guermazi (36) ayant comparé les performances de lecture de plusieurs profils de lecteurs (notamment urgentistes, orthopédistes, rhumatologues, radiologues), a permis de montré une Sp avec IA non inférieure à la Sp sans IA (+5.0% (95% CI: 12.0, 18.0; P = .001 pour la non infériorité). Des études supplémentaires sur un ensemble de données plus grand et plus diversifié seraient nécessaires pour valider nos résultats. Ce jeu de données déséquilibré en faveur des lésions, pourrait aussi s'expliquer par le fait qu'après avoir diagnostiqué une lésion sur la radiographie, un avis d'un chirurgien orthopédiste aurai préconisé un TDM pour préciser la lésion et décider d'une prise en charge (orthopédique ou chirurgicale). Malheureusement, nous ne savions pas quels cas a bénéficié d'un avis orthopédique préconisant un TDM pour statuer non pas sur le diagnostic mais sur la prise en charge.

Dans notre étude, les intervalles de confiance des valeurs prédictive positive avec et sans IA se chevauchaient indiquant que les différences mesurées étaient incertaines et peut être uniquement dues à la taille et au déséquilibre du jeu de données. Par conséquent, il n'y a pas suffisamment de preuves pour conclure que l'IA a un impact réel sur la VPP à l'échelle du cas dans notre étude. Cette observation suggère que l'IA n'affecte pas négativement la précision globale des résultats positifs à ce niveau. Mais là encore, des études futures avec un

échantillon plus large seraient nécessaires pour confirmer nos résultats. Rappelons que l'étude de L. Canoni-Meynet (35) avait permis de montrer une augmentation de la VPP et celle de Bachmann en 2023 (37) avait quant à elle permis de montrer une diminution de 21% des faux positifs avec l'aide de l'IA pour détecter des fractures du squelette appendiculaire, chez des lecteurs non experts.

Notre étude a inclus un faible nombre de lecteurs, avec seulement 30 lecteurs inclus et un nombre très limité de séniors (20% de l'échantillon). Initialement nous avions prévu une analyse en 3 sous-groupes d'expérience : interne, DJ et praticiens. Malheureusement devant le faible nombre de volontaire DJ et praticiens, nous avons dû les réunir dans un seul groupe « sénior ». Bien que les résultats montrent que l'IA améliore la sensibilité et réduit le temps de lecture, ces conclusions sont principalement valables pour les internes, qui constituaient la majorité de l'échantillon. Par conséquent, les résultats doivent être interprétés avec prudence et ne peuvent pas être directement généralisés aux urgentistes seniors. Des études supplémentaires avec un échantillon plus équilibré en termes d'expérience des lecteurs seraient nécessaires pour confirmer ces résultats.

Il parait être important de préciser que l'annotation des radiographies lors des 2 phases pouvait se faire sur ordinateur ou téléphone portable et que cette dernière était plus longue que sur ordinateur. En effet Encord n'est pas développé pour être correctement utilisé sur téléphone portable dans ce contexte. Les temps de lecture extrêmement long pourraient être expliqué par ce mode d'annotation et donc constitué un biais de mesure. Cependant pour chaque lecteur pris individuellement, le même mode d'annotation a été utilisé au cours des 2 phases. Une autre explication qui pourrait être donné aux temps de lecture plus long, pourrait être que l'on suppose que certains lecteurs aient fait des pauses lors de leurs annotations allongeant ainsi le temps de lecture.

Enfin, une des dernières limites de notre étude est que les lecteurs ont analysés le même jeu de donné lors des 2 phases d'annotation. Certes les phases étaient espacées d'une période appelée « wash-out », d'un mois, le faible effectif de notre jeu de données d'analyse, pouvait constituer un biais de mémorisation.

CONCLUSION

En améliorant la sensibilité et le temps de lecture, l'IA pourrait devenir un outil utile pour améliorer les performances et la rapidité de lecture des radiographies aux urgences, en particulier dans des zones où les anomalies sont plus difficiles à détecter (comme le poignet) ou pour des lecteurs avec moins d'expérience (interne).
On peut supposer ces bénéfices pourraient augmenter le temps clinique (temps passé auprès des patients), accélérer le flux de passage aux urgences tout en réduisant la fatigue des urgentistes

Cependant, pour confirmer nos résultats et répondre à nos suppositions, il serait nécessaire de faire une future étude avec un échantillon de cas et de lecteurs plus grand et moins déséquilibré concernant la positivité des lésions et de l’expérience des lecteurs.

ANNEXE 1 : DÉFINITION DES PERFORMANCES DIAGNOSTIQUES

- La **sensibilité (Se)** est définie comme le ratio de vrais positifs (VP) sur la somme des vrais positifs (VP) et des faux négatifs (FN)

$$Sensibilité = \frac{Vrais\ positifs\ (VP)}{VP + Faux\ négatifs\ (FN)}$$

- La **spécificité (Sp)** est définie comme le ratio de vrais négatifs (VN) sur la somme des vrais négatifs (VN) et des faux positifs (FP)

$$Spécificité = \frac{Vrais\ négatifs\ (VN)}{VN + Faux\ positifs\ (FP)}$$

- La **valeur prédictive positive (VPP)** est définie comme le ratio de vrais positif (VP) sur la somme des vrais positifs (VP) et des faux positifs (FP)

$$Valeur\ prédictive\ positive = \frac{Vrais\ positifs\ (VP)}{VP + Faux\ positifs\ (FP)}$$

- La **valeur prédictive négative (VPN)** est définie comme le ratio de vrais négatifs (VN) sur la somme des vrais négatifs (VN) et des faux négatifs (FN)

$$Valeur\ prédictive\ négative = \frac{Vrais\ négatifs\ (VN)}{VN + Faux\ négatifs\ (FN)}$$

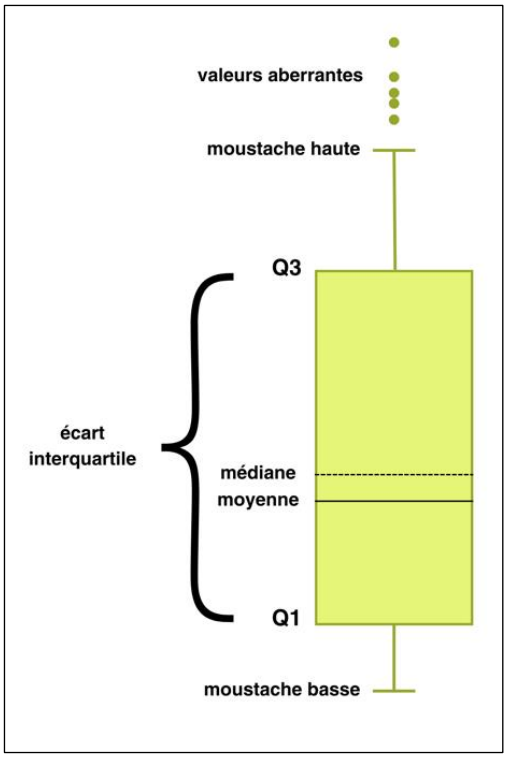
- Le **temps de lecture** correspond au temps passé par le lecteur depuis l'ouverture du dossier jusqu'à sa soumission sur la plateforme Encord.
- **Intervalle de confiance (IC**) : concept statistique utilisé pour estimer la précision d'une mesure, à partir d'un échantillon de données. Plus précisément, un intervalle de confiance fournit une gamme de valeurs, calculée à partir des données échantillonnées, dans laquelle on s'attend à ce que la vraie valeur du paramètre de la population se trouve avec un certain niveau de confiance. Plus l'IC sera étroite, plus l'estimation sera précise et inversement. Dans notre étude, nous avions des IC à 95%, c'est-à-dire que nous étions surs à 95% que la vraie valeur se situe dans l'intervalle.
- **Box plot ou boite à moustache** (Figure 18) : représentation graphique permettant de représenter la distribution des données pour une variable continue.

La ligne pleine indique la médiane où 50% des données sont au-dessus et 50% en dessous, la ligne en pointillée indique la moyenne. Le bas de la boite représente le 1er quartile (Q1) et le haut le 2ème quartile (Q3) correspondant respectivement à 25 et 75% des données. La hauteur de la boite est la différence entre ces 2 quartiles et est appelé intervalle interquartile ou IQR pour interquartile range. 50% des valeurs sont dans la boite.

Les lignes qui s'étendent de la boite sont appelés les moustaches hautes (upper whisker) ou basse (lower whisker), et sont égales respectivement à Q1 + (1,5x IQR) et Q3 + (1,5 x IQR). Les valeurs en dessous ou en dessus des moustaches sont appelés valeurs aberrantes ou outliers.

- **Échelle du cas** : si une anomalie est présente ou non dans un cas
- **Échelle de la lésion** : si chaque lésion individuelle dans une radiographie est correctement identifiée.

Figure 18 : Légende d'une boite à moustache ou box plot



ANNEXE 2 : DONNÉES RELATIVES AUX PASSAGES AUX URGENCES DU CH DE MAUBEUGE

Tableau 17 : Résumé des passages aux urgences du CHM et des examens complémentaires prescrits de 2020 à 2023

Année	Entrées aux urgences adultes en effectif	Dont traumatismes selon liste fournie* en effectif	Soit en %		Dont ayant eu un scanner	Dont ayant eu une radio	Dont ayant eu scanner ou radio
2020	31 541	6 090	19,31%		3,73%	22,99%	25,68%
2021	33 275	6 948	20,88%		4,25%	73,56%	74,61%
2022	35 578	7 078	19,89%		4,31%	73,91%	74,98%
2023	35 955	7 156	19,90%		4,49%	74,06%	75,08%

Tableau 18 : Pourcentage (%) de patients ayant consulté aux urgences du CHM pour traumatismes et ayant eu un compte rendu (CR) d'imagerie à J0 ou J1 de 2020 à 2023

	Nb cas	Avec CR	%	Avec CR Jo	%	Avec CR J1	%
2020	6 090	1 053	17,3%	170	16,1%	76	7,2%
2021	6 948	1 416	20,4%	263	18,6%	198	14,0%
2022	7 078	3 674	51,9%	579	15,8%	813	22,1%
2023	7 156	3 243	45,3%	648	20,0%	1 003	30,9%

ANNEXE 3 : DONNÉES DÉMOGRAPHIQUES SUR LE JEU DE DONNÉES INITIAL (AVANT EXCLUSION)

Tableau 19 : Caractéristiques de la population initiale

	Age *(années)*						Effectif *(nb)*	Sex ratio H/F
	Moyenne	Médiane	1er quartile	3ème quartile	Min	Max	Total	
Population globale	56,2	65,5	36,3	74,5	17	96	62	0,8
Homme	45,5	39	28,5	66,5	17	87	27	
Femme	64,4	70	60	77,5	18	96	35	
≥ 65 ans	76,1	74	68	84	65	96	32	0,25
< 65 ans	34,9	35,5	23,3	41	17	63	30	1,5
Hanche	72,1	76,5	66,8	88,3	24	90	8	1
Cheville	51,6	45	36	71	19	92	17	0,6
Coude	59,6	68	37	73	18	89	17	0,7
Poignet	50,7	48	33	69,8	17	96	20	1
Note : nb signifie nombre								

Tableau 20 : Caractéristique du ground-truth du jeu de données initial

	TDM *(nb)*		Latéralité *(nb)*		
	Positif	Négatif	Droit	Gauche	Droit et Gauche
Population générale	55	5	31	27	4
Homme	23	4	16	10	1
Femme	32	3	15	17	3
Hanche	8	0	2	5	1
Cheville	16	0	10	7	0
Coude	16	1	11	6	0
Poignet	14	6	8	9	3
Note : nb signifie nombre					

BIBLIOGRAPHIE

1. Se questionner : qui a peur de l'IA ? [Internet]. Class'Code IAI; 2020 [cité 13 juill 2024]. (L'Intelligence Artificielle... avec intelligence !; vol. 1). Disponible sur: https://pixees.fr/une-formation-a-lintelligence-artificielle-intelligente/

2. Turing. Computing Machinery and Intelligence. Mind. oct 1950;59:433-60.

3. McCarthy J, Minsky M. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Mag. Aout 1955;27(4).

4. McCarthy J. Recursive functions of symbolic expressions and their computation by machine, Part I. Commun ACM. avr 1960;3(4):184-95.

5. J. J Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proc NatL Acad Sci USA. avr 1982;79,(8):2554-2558,.

6. Y. LeCun, O. Matan, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, et al. Handwritten Zip Code Recognition with Multilayer Networks. Proc Int Conf Pattern Recognit. 1990;(II):35-40.

7. Smith C, McGuire B, Huang T, Yang G. The History of Artificial Intelligence. Univ Washigton. 2006;

8. Moore GE. Cramming more components onto integrated circuits (Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff). IEEE Solid-State Circuits Soc Newsl. sept 2006;11(3):33-5.

9. Tristant Gaudiaut. Statista Daily Data. 2021 [cité 24 août 2024]. Infographie: Le Big Bang du Big Data. Disponible sur: https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde

10. Claire Jenik. Statista Daily Data. 2021 [cité 24 août 2024]. Infographie: Une minute sur Internet en 2021. Disponible sur: https://fr.statista.com/infographie/25402/nombre-de-donnees-generees-sur-internet-par-minute

11. Deng J, Dong W, Socher R, Li LJ, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009 Conf Comput Vis Pattern Recognit. juin 2009;248-55.

12. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386-408.

13. Shortliffe EH. Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. Proc Annu Symp Comput Appl Med Care. 7 oct 1977;66-9.

14. Milvue. TechCare Alert [Internet]. Milvue. Disponible sur: https://www.milvue.com/solutions/techcare-alert/

15. Valérie Bousson, Grégoire Attané, Nicolas Benoist. Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms. Acad Radiol. juill 2023;30(10):2118-39.

16. DREES - 2024 - Les établissements de santé en 2022.pdf [Internet]. [cité 2 août 2024]. Disponible sur: https://drees.solidarites-sante.gouv.fr/sites/default/files/2024-07/ES24%20-%20Fiche%2023%20-%20La%20m%C3%A9decine%20d%E2%80%99urgence.pdf

17. Elvire Demoly, Wilfried Bara, Diane Naouri, Bénédicte Boisguerin, Carla Bianchi. Urgences hospitalières en 2023 : quelles organisations pour la prise en charge des patients ? DREES [Internet]. juill 2024;Etudes et Résultats(1305). Disponible sur: https://drees.solidarites-sante.gouv.fr/sites/default/files/2024-07/ER1305_0.pdf

18. Fabrice Lenglart. DREES. 2023. Démographie des professionnels de santé. Disponible sur: https://drees.shinyapps.io/demographie-ps/

19. Pr Pateron, Pr Gauvrit. L'accès à l'imagerie au cours des urgences : analyse de l'enquête nationale - JFR 2014 [Internet]. Caducee.net. 2014. Disponible sur: https://www.caducee.net/actualite-medicale/12669/l-acces-a-l-imagerie-au-cours-des-urgences-analyse-de-l-enquete-nationale.html

20. Bénédicte Boisguerin, Gwennaëlle Brilhault, Layla Ricroch, Hélène Valdelièvre, Albert Vuagnat. Structures des urgences hospitalières: premiers résultats de l'enquête nationale réalisée par la DREES. DREES. 2014;31-47.

21. Journal officiel. Arrêté du 21 avril 2017 relatif aux connaissances, aux compétences et aux maquettes de formation des diplômes d'études spécialisées et fixant la liste de ces diplômes et des options et formations spécialisées transversales du troisième cycle des études de médecine [Internet]. Sect. Annexe II-II avr 21, 2017 p. Maquette 13. Disponible sur: https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000047115727

22. Newman-Toker DE, Peterson SM, Badihian S, Hassoon A, Nassery N, Parizadeh D, et al. Diagnostic Errors in the Emergency Department: A Systematic Review [Internet]. Agency for Healthcare Research and Quality (AHRQ); 2022 déc. Disponible sur: https://effectivehealthcare.ahrq.gov/products/diagnostic-errors-emergency/research

23. FEDORU. Panorama des ORU [Internet]. 2022. Disponible sur: file:///Users/ilonabellotto/Downloads/PANORAMA_FEDORU_2022_WEB.pdf

24. Hanna TN, Lamoureux C, Krupinski EA, Weber S, Johnson JO. Effect of Shift, Schedule, and Volume on Interpretive Accuracy: A Retrospective Analysis of 2.9 Million Radiologic Examinations. Radiology. avr 2018;287(1):205-12.

25. Jacques T, Cardot N, Ventre J, Demondion X, Cotten A. Commercially-available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray, compared to a CT-based ground truth. Eur Radiol. 3 nov 2023;34(5):2885-94.

26. Parpaleix A, Parsy C, Cordari M, Mejdoubi M. Assessment of a combined musculoskeletal and chest deep learning-based detection solution in an emergency setting. Eur J Radiol Open. 10 mars 2023;10:100482.

27. 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. juill 2021;300(1):120-9.

28. Jones RM, Sharma A, Hotchkiss R, Sperling JW, Hamburger J, Ledig C, et al. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. Npj Digit Med. 30 oct 2020;3(1):144.

29. Fu T, Viswanathan V, Attia A, Zerbib-Attal E, Kosaraju V, Barger R, et al. Assessing the Potential of a Deep Learning Tool to Improve Fracture Detection by Radiologists and Emergency Physicians on Extremity Radiographs. Acad Radiol. 1 mai 2024;31(5):1989-99.

30. Shin HJ, Han K, Ryu L, Kim EK. The impact of artificial intelligence on the reading times of radiologists for chest radiographs. Npj Digit Med. 29 avr 2023;6(1):82.

31. Laur O, Wang B. Musculoskeletal trauma and artificial intelligence: current trends and projections. Skeletal Radiol. févr 2022;51(2):257-69.

32. 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. juin 2019;100(6):327-36.

33. Dreizin D, Staziaki PV, Khatri GD, Beckmann NM, Feng Z, Liang Y, et al. Artificial intelligence CAD tools in trauma imaging: a scoping review from the American Society of Emergency Radiology (ASER) AI/ML Expert Panel. Emerg Radiol. 14 mars 2023;30(3):251-65.

34. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol. 14 avr 2022;32(11):7998-8007.

35. Canoni-Meynet L, Verdot P, Danner A, Calame P, Aubry S. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. Diagn Interv Imaging. déc 2022;103(12):594-600.

36. Guermazi A, Tannoury C, Kompel AJ, Murakami AM, Ducarouge A, Gillibert A, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. Radiology. mars 2022;302(3):627-36.

37. Bachmann R, Gunes G, Hangaard S, Nexmann A, Lisouski P, Boesen M, et al. Improving traumatic fracture detection on radiographs with artificial intelligence support: a multi-reader study. BJR|Open. 12 déc 2023;6(1).

AUTEURE : Nom : Bellotto **Prénom : Ilona**

Date de soutenance : 17 octobre 2024

Titre de la thèse : Intérêt pour les urgentistes de la préanalyse des radiographies traumatiques par l'IA

Thèse - Médecine - Lille « 2024 »

Cadre de classement : *Radiologie*

DES + FST/option : *Médecine d'urgence*

Mots-clés : intelligence artificielle, machine learning, radiologie, urgences, fracture, luxation, formation

Résumé : ***Contexte*** : l'augmentation des passages aux urgences et le manque de médecins urgentistes et radiologues a conduit à une surcharge de travail. Aux urgences, l'interprétation des radiographies est souvent réalisée par des urgentistes non formés spécifiquement, augmentant ainsi le risque d'erreurs, notamment pour les fractures. Le développement de l'intelligence artificielle (IA), comme l'outil Milvue Suite, vise à améliorer la qualité et l'efficacité du diagnostic dans ce contexte. ***Objectifs*** : l'objectif principal de cette étude était d'évaluer l'impact de l'utilisation de Milvue Suite sur les performances diagnostiques des urgentistes exerçant dans la lecture des radiographies traumatiques. L'étude visait à déterminer si l'IA améliore la sensibilité et la spécificité des urgentistes dans la détection des lésions traumatiques. Les objectifs secondaires étaient d'évaluer l'effet de l'IA sur le temps de lecture des radiographies et la comparaison des performances entre internes et séniors, avec et sans IA. ***Matériels et Méthodes*** : étude rétrospective monocentrique ayant inclus 49 cas de radiographies traumatiques d'adultes, associées à des TDM comme référence diagnostique. Les radiographies étaient annotées en deux phases : une première sans IA, puis une deuxième avec l'aide de Milvue Suite. 30 lecteurs (24 internes et 6 séniors) ont participé à cette étude. Les performances diagnostiques (Se, Sp, VPP et VPN) et le temps de lecture ont été comparés entre les deux phases. ***Résultats*** : L'IA Milvue Suite a amélioré la Se des lecteurs, avec une augmentation globale de 68 % à 88 %. L'amélioration des performances diagnostiques était plus marquée chez les séniors, avec un gain de sensibilité de 23%, contre 18% chez les internes. En revanche, il n'a pas été possible d'interpréter la Sp et VPN en raison du faible nombre de cas négatifs dans l'échantillon. De plus, l'IA a permis de réduire significativement le temps médian de lecture, passant de 88,95s sans IA à 48,08s avec IA, représentant une réduction de 40,87s. ***Conclusion*** : L'IA Milvue Suite a amélioré la Se et réduit le temps de lecture des radiographies traumatiques par les urgentistes, à la fois chez les séniors et les internes. Cependant, le faible effectif de cas constitue une limite de l'étude. L'IA apparaît comme un outil prometteur pour soutenir les urgentistes.

Composition du Jury :

Président :

Monsieur le Professeur Eric WIEL

Assesseurs :

Monsieur le Docteur Jean Marie RENARD
Monsieur le Docteur Jérôme MIZON

Directeur de thèse :

Monsieur le Docteur Romain DEWILDE