Recommandations

Intelligence artificielle en médecine : comment éviter un mésusage

Explications et recommandations pour les praticiens militaires

R. CHAPELLE[(1)], N. IONESCU[(2)], N. DE L'ESCALOPIER[(3)], F. BOMPAIRE[(4)], S. KREMER[(5)], N. VAYATIS[(6)], D. RICARD[(7)]

(1,7,4) Hôpital National d'Instruction des Armées Percy, Service de Neurologie, Clamart
(1,7,4) École du Val-de-Grâce, Paris
(1,7,3,4,6) Centre Borelli, CNRS, SSA, INSERM, Université Paris-Saclay, Université Paris Cité, ENS Paris-Saclay, Gif-sur-Yvette
(2,5) Centre Hospitalier Universitaire de Strasbourg, Service d'imagerie, Strasbourg
(3) Hôpital National d'Instruction des Armées Percy, Service de Chirurgie Orthopédique, Traumatologique et Réparatrice des Membres, Clamart
(5) Centre Hospitalier Universitaire de Strasbourg, Service d'imagerie, Strasbourg

Auteur correspondant : damien.ricard@m4x.org

Open access DOI
10.17184/eac.9439

Reçu
le 29 janvier 2024

Accepté
le 12 mai 2025

Publié
le 11 juin 2025

Conflit d'intérêt
Les auteurs ne déclarent pas de conflit d'intérêt.

Résumé :
Les usages de l'intelligence artificielle se sont multipliés au cours des dernières années, y compris en médecine.
Ils s'accompagnent d'une promesse de performances inégalées à faible coût, ce qui les a rendus particulièrement attrayants aux yeux des non-spécialistes. Ils ne sont pourtant pas exempts de dangers provenant souvent d'une mauvaise compréhension du fonctionnement des algorithmes d'intelligence artificielle.
Par cet article, nous avons voulu sensibiliser les praticiens militaires aux mésusages de l'intelligence artificielle, et leur donner les clés de compréhension utiles pour les éviter.
Nous présentons pour cela de façon didactique certains aspects techniques indispensables pour saisir la nature de l'intelligence artificielle. Nous proposons en parallèle des recommandations relatives à chacun de ces aspects, déclinées en versions qui s'adressent respectivement aux concepteurs de modèles d'apprentissage automatique et aux praticiens qui les utilisent. Enfin, nous illustrons les dangers d'un mésusage de l'intelligence artificielle par des cas représentatifs issus du champ de la neuroradiologie, et discutons des choix organisationnels entre IA et pratique médicale susceptibles de limiter ces mésusages.

Mots-clés : Apprentissage automatique. Intelligence artificielle. Mésusage. Neuroradiologie. Médecine d'armée.

Abstract:
ARTIFICIAL INTELLIGENCE IN MEDICINE, HOW TO AVOID MISUSE: KEYS TO UNDERSTANDING AI AND RECOMMENDATIONS FOR MILITARY PRACTITIONERS
Using artificial intelligence in such fields as medicine has kept increasing in recent years.
They come with the promise of unprecedented performance at low cost, offering appealing prospects to non-specialists. However, it is not without dangers which often stem from a misunderstanding of the behavior of artificial intelligence algorithms.
In this article, we aim to raise awareness among military practitioners about the potential misuses, and provide them with the necessary understanding to avoid pitfalls. In that respect, we present some technical aspects essential to grasping the nature of artificial intelligence.
In parallel, we offer recommendations for each of these aspects, intended respectively for machine learning engineers and for the practitioners who use these models. Finally, we illustrate the dangers of misusing artificial intelligence with representative cases from the field of neuro-radiology, and discuss solutions to limit these misuses through judicious workflow choices.

Keywords: Artificial intelligence. Machine learning. Misuse. Neuroradiology. Military medicine.

Introduction

L'Intelligence artificielle (IA) connaît depuis quelques années une popularité inédite auprès de publics variés, y compris de non-experts. Chacun peut désormais en faire usage : agents conversationnels, véhicules autonomes, maisons intelligentes s'insèrent progressivement dans notre quotidien. Il faut dire que les promesses sont immenses. Elles ont trait principalement à l'anticipation d'évènements et à l'automatisation de tâches, et se déclinent en une infinité de cas d'usage.
En médecine, il peut notamment s'agir de générer des biomarqueurs diagnostiques et pronostiques, d'automatiser l'interprétation d'examens (1) et de faciliter la rédaction de contenu scientifique et clinique (2). La pertinence de ces usages découle notamment de la capacité des modèles d'IA à traiter des volumes importants de données, bien au-delà de la perception et des capacités de traitement simultané de l'intelligence humaine en un temps optimisé. Les méthodes d'exploration de données dites « massives » en sont une illustration frappante, comme c'est le cas pour les bases « omiques ». Parmi leurs applications cliniques, on peut ainsi citer l'identification de biomarqueurs en vue d'une médecine personnalisée (3).
Si les mérites de l'IA sont largement diffusés, il n'en va pas de même des dangers de ses mésusages. Nous les définissons ici comme tout écart non intentionnel aux bonnes pratiques concernant la conception ou l'utilisation d'outils d'IA, présentant un risque de conséquences indésirables.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

Des cas documentés attestent de la réalité et de la gravité potentielle de ces mésusages. Lors de tests préliminaires en 2020, l'agent conversationnel médical conçu par l'entreprise Nabla, reposant sur le modèle GPT-3, a, par exemple, conseillé à un patient fictif de se suicider (4). La prévention de tels écueils représente un enjeu particulièrement important dans les domaines sensibles du fait de la nature des données traitées ou de leurs applications. À cet égard, les usages de l'IA par les médecins militaires, chercheurs ou cliniciens, sont en bonne position, et on peut aisément imaginer la portée qu'aurait un cas documenté de mésusage de l'IA avec des conséquences opérationnelles délétères que ce soit dans la prise en charge de patients ou dans l'expertise médico-militaire et en particulier la sélection de candidats à certains postes ou opérations.

L'expérience montre que de tels écueils sont souvent liés à une mauvaise appréhension de l'IA, notamment sur le plan de ses conditions d'applicabilité et de l'interprétation des résultats qu'elle fournit (5,6). Dans cet article, nous avons donc souhaité fournir aux praticiens militaires les clés de compréhension nécessaires à la prévention de ces mésusages, pour des applications à la fois cliniques et scientifiques. Nous suivons pour cela une approche thématique, déclinée selon des types de mésusages fréquents correspondant à autant d'aspects techniques de l'intelligence artificielle, et illustrée par des exemples concrets issus du champ de la neuroradiologie, mais facilement transposables à la biologie ou l'analyse sémiologique ou épidémiologique.

1 Intelligence artificielle, apprentissage automatique, modèles génératifs : de quoi parle-t-on ?

On peut définir l'IA comme le domaine de l'informatique portant sur les systèmes capables de réaliser des tâches réputées mettre en jeu des capacités associées à l'intelligence humaine (7), comme la résolution de problèmes complexes ou la reconnaissance de motifs (8). Diverses approches ont été proposées historiquement pour construire de tels systèmes. Certaines reposent sur l'alimentation de bases de connaissances, ensuite traitées à l'aide de règles d'inférence pour résoudre un type de problèmes. On parle notamment de systèmes experts (9).

Une autre approche, bien plus populaire actuellement (et à laquelle nous limitons donc notre propos dans la suite de cet article), est celle de l'apprentissage automatique. Elle repose sur des modèles[1] dont le comportement n'est pas défini par une base de connaissances constituée de règles dictées manuellement, mais par des bases de données brutes, souvent enrichies par des annotations manuelles qui sont intégrées par le biais d'un processus appelé entraînement. Ces modèles sont alors capables de réaliser des prédictions lorsqu'on leur présente de nouvelles données numérisées et structurées de manière similaire aux données d'entraînement (12). Lorsque les données sont enrichies par des annotations manuelles (appelées aussi *labels* ou étiquettes), on parle d'apprentissage supervisé. Pour illustrer ce cas, on peut imaginer un problème de classification visant à prédire, chez des patients bactériémiques hospitalisés, la bactérie en cause chez chacun d'eux. Une approche d'apprentissage supervisé consisterait

[1] Ceux-ci ne sont, en général, pas des modèles au sens statistique du terme. En effet, un modèle statistique est une famille de lois de probabilité dont on fait l'hypothèse qu'elle contient la loi dont proviennent des données observées (10). Ici, nous employons le terme « modèle » dans son sens informel en apprentissage automatique, qui désigne la règle prédictive qui résulte d'un processus d'apprentissage (11).

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

à entraîner pour cela un modèle d'apprentissage automatique à l'aide d'un jeu de données (si tant est qu'il en existe en version totalement numérisée) contenant des variables cliniques et biologiques mesurées chez un grand nombre de patients hospitalisés bactériémiques, parmi lesquelles l'espèce bactérienne en cause chez chacun d'eux. En présence d'une nouvelle ligne de données cliniques et biologiques, le modèle pourrait alors prédire l'espèce bactérienne la plus probablement en cause, compte tenu des données qui ont servi à l'entraîner. Si la classe de modèles de classification choisie est probabiliste, un tel modèle peut également fournir une probabilité estimée que ses prédictions soient exactes (ce qui n'est pas sans poser de problèmes de fiabilité en haute dimension) (13). Lorsque les données brutes ne sont pas enrichies par des annotations manuelles, on parle d'apprentissage non supervisé. Les applications typiques de l'apprentissage non supervisé sont le partitionnement d'observations en clusters, la détection d'anomalies ou de nouveautés, la réduction de dimension, et l'estimation de densité. Nous ne les abordons pas directement dans la suite de cet article, car les difficultés théoriques et techniques qu'elles soulèvent leur sont, pour une part importante, spécifiques. En particulier, on peut mentionner la difficulté de valider des méthodes d'apprentissage non supervisé en l'absence d'étiquettes alors qu'en apprentissage supervisé, ce sont justement ces étiquettes qui fournissent un critère de succès. À défaut de tels critères, la validation des algorithmes d'apprentissage non supervisé peut reposer sur des dires d'expert et comporte une part inévitable d'arbitraire (12).

Il existe une grande diversité de classes de modèles d'apprentissage automatique.
À l'extrême, des outils comme la régression linéaire peuvent être considérés comme tels lorsqu'ils sont utilisés pour des tâches de prédiction plutôt que d'inférence statistique. D'autres modèles sont adaptés au traitement de données structurées de façon non tabulaire, par exemple des images, du texte ou des signaux physiologiques.
Une approche qui s'est particulièrement distinguée pour les images, la vidéo, le texte, et l'audio, est celle de l'apprentissage profond, qui a recours à des réseaux de neurones artificiels organisés en de nombreuses couches, dont les connexions et transformations non linéaires permettent d'extraire, à chaque niveau plus profond, des caractéristiques ou représentations des données de plus en plus abstraites et complexes (14).

Lorsqu'un modèle fournit des prédictions stochastiques (c'est-à-dire qu'elles ne correspondent pas à une espérance conditionnelle, mais à un tirage aléatoire *via* une loi de probabilité), on peut le qualifier de modèle génératif. L'appellation est particulièrement employée lorsque l'espace des prédictions est complexe : images, vidéos, etc. Les grands modèles de langage en sont un exemple typique : GPT, Gemini, Mistral en font partie (15). Ils reposent sur des techniques d'apprentissage profond pour « générer » (en fait sous la forme d'un tirage depuis une loi de probabilité conditionnelle) des séquences de mots (ou « jetons », en anglais *tokens*) étant données les séquences précédentes. Ce sont ces modèles qui ont contribué à la popularité de l'IA auprès du grand public, au point qu'aujourd'hui le terme même d'IA y soit fortement associé. Cela renvoie au lien philosophique étroit qui existe entre pensée et langage, et qui a inspiré des critères historiques d'intelligence artificielle : une machine qui parle est volontiers perçue comme une machine intelligente, comme illustré par le test de Turing (16,17). Ces systèmes ne peuvent être développés que dans des domaines pour lesquels existent des bases de données numérisées de très grande ampleur, telles que celles issues de l'activité sur Internet de parts importantes de la population mondiale.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

Ce type de modèles n'est pas adapté aux domaines pour lesquels les bases de données numérisées, même si elles sont parfaitement classées et annotées, ne dépassent pas une certaine taille. Il a été décliné en versions ciblant des usages médicaux (18), parmi lesquelles Meditron (19) et MedPaLM (20).

2 Connaître et prévenir les mésusages de l'IA

À la lumière des définitions données ci-dessus, cette section propose une description des principaux cas possibles de mésusage de l'apprentissage automatique par les praticiens militaires, distingués selon les aspects techniques qu'ils concernent. Pour chacun de ces cas, des recommandations sont proposées aux chercheurs et cliniciens afin de favoriser l'adoption de mesures de prévention adéquates.
Au sein de cette section, nous prenons plusieurs fois comme exemple un problème d'apprentissage supervisé typique des applications de l'IA en médecine (21) : celui de la classification des patients bactériémiques hospitalisés selon la bactérie en cause. Nous nommons BACT l'étude fictive dont les données servant à entraîner le modèle de classification seraient issues.

2.1 Données d'entraînement

2.1.1 Contexte

Comme mentionné dans la section précédente, les algorithmes d'apprentissage automatique peuvent être vus comme des méthodes d'inférence statistique particulièrement adaptées au cas de données complexes, éventuellement structurées. On considère alors qu'ils visent à estimer les propriétés d'une loi de probabilité à partir d'un échantillon tiré de celle-ci. Par exemple, dans le cas d'un problème de régression, la propriété d'intérêt est une espérance conditionnelle. Dans le cas d'un problème d'apprentissage non supervisé, il peut s'agir des modes de la loi. La théorie de la décision fournit le cadre formel permettant d'établir sous quel sens, et à quelles conditions, les estimations fournies par une classe de modèles sont les meilleures (22).
Elle repose sur l'hypothèse indispensable que les données d'entraînement sont un tirage aléatoire depuis la population dont on cherche à estimer les caractéristiques. Lorsque cette hypothèse n'est pas vérifiée, on parle de biais d'échantillonnage, et on dit parfois que les données d'entraînement sont biaisées[2] ou présentent un biais d'échantillonnage (23). Il existe plusieurs sources de biais pour des données d'entraînement. Le biais peut être introduit lors de la constitution de la base de données.
Dans l'étude BACT, cela surviendrait, par exemple, si on proposait plus volontiers aux hommes qu'aux femmes d'être inclus. Dans ce cas, l'échantillon constitué ne pourrait pas être considéré comme provenant d'un tirage aléatoire depuis la population bactériémique dans son ensemble. Notons que même en l'absence de biais, la qualité des données d'entraînement d'un modèle peut être faible, notamment du fait d'erreurs de mesure (à cause de la précision des appareils utilisés ou d'erreurs humaines de recueil des données, par exemple). Une autre situation fréquente est celle d'erreurs d'annotation, lorsque les étiquettes sont fournies par des experts et que la qualité de l'annotation est imparfaite (24). De telles données sont dites bruitées (25).

[2]Notons qu'il s'agit d'un abus de langage, car, au sens strictement statistique du terme, seuls des estimateurs peuvent être biaisés, et non des données. Cet abus de langage signifie en fait que le caractère informatif de l'échantillonnage n'est pas pris en compte par l'estimateur considéré.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

2.1.2 Description du mésusage

L'utilisation de données de qualité insuffisante pour entraîner un modèle (qu'elles soient biaisées ou bruitées), sans mesures de correction adéquates, constitue un mésusage. Il en va de même de l'utilisation d'un modèle au sein d'une population trop différente de celle sur laquelle il a été entraîné. Dans les deux cas, le risque est d'abord celui d'une mauvaise performance du modèle. Lorsque les relations observées entre des variables sont altérées par du bruit, ou par un biais, elles conduisent à une mauvaise estimation des paramètres du modèle, et donc à de moins bonnes prédictions.
Dans le cas de données d'entraînement biaisées, la performance du modèle tend à se dégrader encore plus pour les individus qui y sont sous-représentés (par exemple, les femmes si celles-ci sont moins volontiers incluses dans l'étude BACT). C'est ainsi que des biais statistiques peuvent reproduire des biais sociaux, posant des questions éthiques lors de la constitution de bases de données qui sont de plus en plus abordées dans la littérature. Le cadre technique est alors celui de l'équité (*fairness*), qui a donné lieu au développement de métriques et mécanismes de correction spécifiques (26).

2.1.3 Recommandations pour le chercheur

Pour limiter ces risques lorsqu'on conçoit un modèle d'apprentissage automatique, la clé est de travailler avec des données de bonne qualité. Il y a pour cela des critères méthodologiques à prendre en compte dès la conception des études dont proviennent ces données : simplification de leur recueil et de leurs flux, optimisation du contrôle qualité, etc. Pour se prémunir des biais, il est indispensable de définir précisément la population sous-jacente, et de s'assurer que tous les individus ainsi désignés ont autant de chance d'être inclus dans la base de données. Cela peut reposer sur des stratégies analogues à celles utilisées pour la prévention des biais de sélection dans les études épidémiologiques : formation des investigateurs, campagnes informatisées de recrutement de participants, etc. Lorsqu'on suspecte l'existence de données biaisées ou bruitées malgré ces mesures, il est conseillé de recourir à des algorithmes d'apprentissage automatique suffisamment robustes. Selon la classe de modèles considérée, il peut s'agir, par exemple, de chercher à maximiser la marge (c'est-à-dire la distance de l'hypersurface séparatrice à l'observation du jeu d'entraînement la plus proche) (27), ou d'utiliser une fonction de coût absolu plutôt que quadratique. En cas de préoccupations concernant la performance du modèle dans certaines sous-populations spécifiques, les outils du champ de l'équité peuvent être utilisés. En plus de métriques permettant de quantifier le phénomène, ils incluent des méthodes de correction reposant par exemple sur des fonctions de perte surpondérées pour les observations appartenant à des groupes minoritaires. Ces approches peuvent être utilisées même en l'absence de biais, car un modèle peut s'avérer peu performant dans une sous-population simplement, car cette sous-population est de petite taille, et donc peu représentée dans les données d'entraînement. L'augmentation des données d'entraînement à l'aide d'observations synthétiques (générées de sorte à reproduire les propriétés statistiques d'intérêt obtenues à partir des observations originales, c'est-à-dire les données mesurées) est alors une option envisageable pour limiter ce phénomène (28).
Enfin, notons que certains types de modèles, particulièrement les réseaux de neurones, peuvent bénéficier d'un apprentissage par transfert permettant de les adapter à un autre type de tâche que celle pour laquelle ils ont été entraînés initialement (on parle

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

aussi de *fine tuning*) (29). Ces approches peuvent, à moindre coût, limiter la perte de performance lorsqu'un modèle préentraîné est utilisé dans un cadre qui ne correspond pas exactement à celui de son entraînement initial (par exemple dans une sous-population).

2.1.4 Recommandations pour le clinicien

En matière de données d'entraînement, le clinicien a deux rôles majeurs à jouer : d'une part, dans la constitution des jeux de données, d'autre part dans l'évaluation de leur qualité. En effet, il est souvent le fournisseur des données, soit qu'il les ait recueillies dans leur ensemble, soit qu'il ait annoté en tant qu'expert un jeu de données existant. Sa responsabilité est alors d'introduire le moins possible de bruit et, le cas échéant, de biais, dans ces données. Pour cela, un critère essentiel est qu'il possède un niveau de compétence suffisant pour une telle tâche[3]. Outre cela, les aspects organisationnels évoqués ci-dessus sont primordiaux pour cela. Idéalement, ils doivent permettre de modifier la structure du jeu de données si celle-ci s'avère inadaptée à la tâche d'apprentissage automatique envisagée, ou si elle ne correspond pas à la sémantique du domaine considéré. Par exemple, dans l'étude BACT, si un nombre significatif de patients inclus sont co-infectés par plusieurs espèces bactériennes, alors la structure de la base de données doit permettre d'attribuer à chaque patient autant d'étiquettes que d'espèces bactériennes en cause chez lui. Le problème d'apprentissage automatique devient alors un problème classification multiclasses. Si l'on choisit de restreindre la saisie à une unique espèce bactérienne par patient, certains cliniciens risquent par exemple de saisir l'espèce avec le délai de pousse le plus court, tandis que d'autres pourraient saisir celle réputée la plus pathogène. Les différentes interprétations possibles de la variable « espèce bactérienne » sont alors source de bruit. Enfin, la principale recommandation pour les utilisateurs de modèles d'apprentissage automatique pré-entraînés est de vérifier que les populations qui ont servi à entraîner ces modèles sont bien représentatives de celles pour lesquelles on souhaite les utiliser. Dans le cas contraire, il faut s'attendre à des performances diminuées par rapport à celles estimées par les auteurs du modèle, particulièrement pour les sous-groupes minoritaires.

2.2 Surajustement et sous-ajustement

2.2.1 Contexte

En apprentissage automatique, on cherche à prédire la valeur d'une variable au sein d'une population cible à partir des valeurs prises par d'autres variables. Par hypothèse, cette population cible n'est toutefois pas accessible dans son intégralité, et les modèles d'apprentissage automatique sont entraînés à partir d'un échantillon de celle-ci. Les fluctuations d'échantillonnage peuvent alors conférer à l'échantillon des caractéristiques particulières, qui diffèrent en partie de celles de la population globale dont l'échantillon est issu. Par exemple, dans l'étude BACT, il est possible que l'échantillon constitué contienne par hasard une forte proportion de femmes parmi les

[3]Notons que cette question de l'origine de l'annotation est centrale, non seulement pour assurer sa qualité, mais aussi pour augmenter le niveau de confiance des utilisateurs vis-à-vis d'un modèle d'IA.
Il y a là un enjeu de transparence auquel il faudrait idéalement répondre en publiant l'origine des données et de leur annotation, et même, si possible, en publiant les données elles-mêmes.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

patients ayant une bactériémie à staphylocoque. En même temps, il est possible qu'un grand nombre de ces patients soient atteints d'une infection sur cathéter, parce qu'il existe une association réelle entre ces deux variables dans la population sous-jacente. La première association observée dans l'échantillon (staphylocoque *vs* sexe féminin) serait principalement due aux fluctuations d'échantillonnage, tandis que la seconde (staphylocoque *vs* infection sur cathéter) refléterait une association existante dans la population sous-jacente. Le problème est qu'il n'est pas possible de distinguer le premier type d'association du second à partir de l'échantillon seul. Par conséquent, un modèle d'apprentissage automatique qui serait trop fidèle à ses données d'entraînement risquerait de présenter des performances médiocres lorsqu'appliqué à un autre échantillon issu de la même population sous-jacente, et ce malgré des performances excellentes sur ses données d'entraînement. On parle alors de surajustement ou sur-apprentissage (figure 1). Le sous-ajustement est défini de façon analogue, et survient lorsqu'un modèle ne capture pas suffisamment les caractéristiques observées de ses données d'entraînement.

2.2.2 Description du mésusage

Le phénomène de surajustement donne lieu à deux principaux types de mésusages. D'une part, le surajustement conduit à une diminution des performances du modèle, ce qui est en soi indésirable : on parle d'erreur de généralisation. Il en va de même du sous-ajustement. Mais, surtout, la publication d'un modèle surajusté peut être trompeuse quant à cette performance si les seules données utilisées pour l'évaluer sont celles d'entraînement. Dans ce cas, on peut surestimer considérablement les performances réelles du modèle.

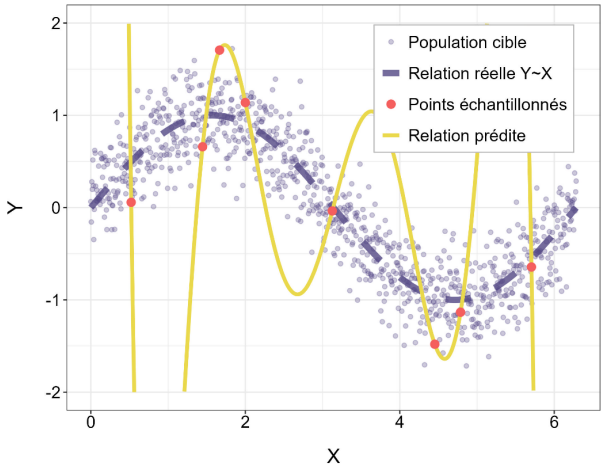


Fig. 1 : Exemple de situation de surajustement. Les données d'entraînement (points rouges) sont choisies au hasard au sein d'une population cible (points bleus). Un modèle de régression polynomiale est construit à partir de ces données (courbe jaune). Ses performances sont excellentes sur les données d'entraînement : la courbe passe par tous les points échantillonnés. Pourtant, la relation modélisée s'écarte nettement de la relation réelle dans la population sous-jacente (courbe bleue).

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

2.2.3 Recommandation pour le chercheur

Le surajustement a été qualifié de « malédiction de l'apprentissage automatique » (30), et il existe une littérature abondante sur les façons de le combattre. Les méthodes proposées visent à garantir une parcimonie quant à la complexité du modèle de façon à trouver un équilibre entre sous- et surajustement. On parle, au sens large, de régularisation. Les hyperparamètres du modèle peuvent être utilisés pour cela. Il s'agit des paramètres dont la valeur contrôle le comportement du modèle en phase d'apprentissage, par exemple le nombre d'itérations pour les modèles entraînés par descente de gradient (c'est-à-dire en suivant la direction dans l'espace des paramètres associée à la plus forte pente négative pour la fonction de perte considérée), la profondeur maximale des arbres de décision, le nombre d'apprenants faibles d'un modèle boosté. S'ils ne le sont pas manuellement, ces hyperparamètres peuvent être déterminés à partir d'algorithmes d'optimisation reposant sur des données d'entraînement et de validation (31,32). De tels algorithmes visent généralement à minimiser l'erreur de généralisation. L'estimation de cette erreur de généralisation, pour un modèle déjà ajusté, est un problème connexe qui est central dans le deuxième type de mésusage décrit ci-dessus. En cas d'abondance de données disponibles, l'approche proposée est de les découper (aléatoirement) en données d'entraînement, de validation et de test, les deux premières étant utilisées pour la sélection du modèle et les dernières pour estimer ses performances. Si le volume de données est insuffisant, il est possible de recourir à des méthodes de réutilisation de données (validation croisée, *bootstrap*) ou analytiques (reposant en particulier sur des mesures de complexité des modèles) (12,33,34).

2.2.4 Recommandation pour le clinicien

Le clinicien confronté à un modèle d'apprentissage automatique doit afficher la plus grande prudence quant aux performances prétendues du modèle. Le critère à rechercher absolument est l'existence d'un échantillon de test utilisé uniquement pour estimer ces performances. Lorsqu'un tel échantillon n'était pas disponible pour la calibration des hyperparamètres, il faut identifier les choix méthodologiques ayant permis de réaliser cette calibration (validation croisée, etc.) et s'assurer de leur validité. Idéalement, les éléments méthodologiques recueillis doivent permettre de réaliser une étude de validation ad hoc (banc d'essai), par exemple à l'aide de jeux de données de référence pour le problème d'apprentissage automatique considéré.
Quelle que soit la méthodologie identifiée, il faut garder à l'esprit que la moindre implication des données de test dans l'entraînement du modèle ou le choix de ses hyperparamètres rend l'estimation de l'erreur de généralisation invalide (avec une tendance générale à la surestimation).

2.3 Explicabilité

2.3.1 Contexte

Dans les années 1970, les systèmes experts comme MYCIN étaient capables de décrire les critères (règles d'inférences, informations fournies par l'utilisateur, etc.) ayant conduit à une réponse donnée. Cela leur conférait une grande explicabilité (ou interprétabilité). L'explicabilité des modèles d'apprentissage automatique n'admet pas de définition générale, mais peut être abordée techniquement (et même philosophiquement) par le biais de concepts distincts.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

Le caractère explicable d'un modèle peut, par exemple, viser à rendre compte de liens causaux. Il peut aussi renvoyer à une notion de fiabilité, ou encore de raisonnement contrefactuel. L'explicabilité d'un modèle se décline par ailleurs selon le public considéré (concepteurs, utilisateurs, acquéreurs, etc.) et son contexte d'utilisation. Enfin, elle peut renvoyer au fonctionnement du modèle dans son ensemble (explicabilité globale), ou au processus ayant conduit à une prédiction en particulier (explicabilité locale) (35).

De façon générale, l'explicabilité est une caractéristique considérée comme souhaitable dans de nombreuses applications de l'apprentissage automatique, pour des raisons notamment éthiques et réglementaires. Elle est d'ailleurs un champ de recherche technique très actif, sous la dénomination de « *XAI* » (« *eXplainable Artificial Intelligence* »). En médecine, par exemple, la possibilité d'expliquer le comportement d'un modèle d'IA peut favoriser son adoption par les praticiens, son acceptation par les patients et sa certification par les autorités sanitaires (36).
En outre, un niveau élevé d'interprétabilité peut permettre de repérer plus facilement les défauts d'un modèle, et de les corriger le cas échéant. Malheureusement, tous les modèles d'apprentissage automatique n'ont pas le même niveau d'explicabilité (pour une conception donnée de l'explicabilité, comme décrit ci-dessus).

En ce qui concerne les règles de décision obtenues par apprentissage, et une notion d'explicabilité causale et globale, les règles considérées comme explicables sont les modèles linéaires, les arbres de décision (figure 2), et les forêts d'arbre (par le biais de l'importance des variables). Par opposition, un réseau de neurones entraîné ne peut classiquement être exprimé que comme un ensemble de coefficients insaisissable par l'esprit humain (37) ; de telles classes de modèles sont décrites comme des « boîtes noires » et sont peu explicables (au sens donné ci-dessus).
Ce constat appelle d'ailleurs une remarque d'ordre sémantique. Si l'IA vise historiquement à mimer l'intelligence humaine, les modèles utilisés actuellement ne peuvent prétendre reposer sur les mécanismes typiquement associés à cette dernière.
En particulier, on ne peut affirmer que ces modèles déroulent des raisonnements, même lorsqu'ils sont entraînés pour des tâches complexes : en fin de compte, ils ne font toujours que produire des valeurs statistiquement probables étant donnés leurs paramètres. Il en va de même des IA génératives : en fait de génération, il ne s'agit que de tirages aléatoires depuis une loi de probabilité. La « génération » de contenu non capturé par de telles lois est impossible.

Notons toutefois que le caractère séquentiel du raisonnement humain peut être reproduit de façon frappante par certains grands modèles de langage, par exemple deepseek-r1. Ce modèle est capable d'exposer une phase de réflexion (résultant d'instructions de type « chaîne de pensée ») en la mettant en contexte avec les échanges précédents, ce qui lui confère, dans un certain sens, un niveau élevé d'explicabilité locale (38).

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

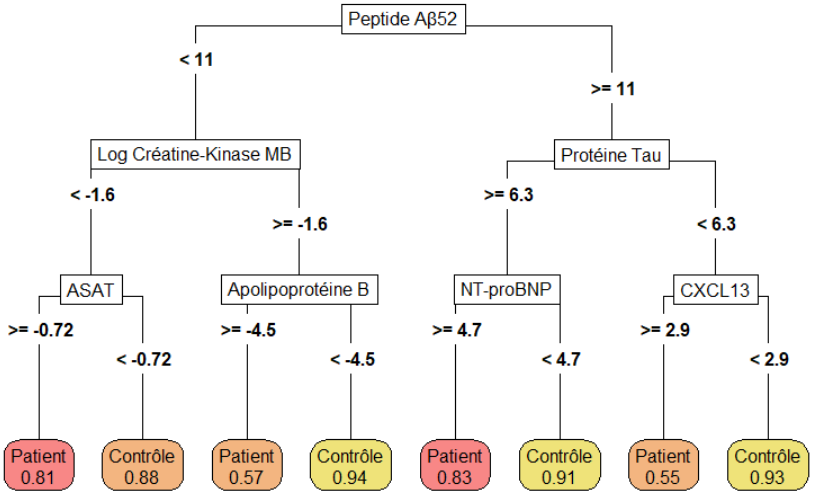


FIG. 2 : Exemple de modèle interprétable : un arbre de classification. L'arbre est entraîné à prédire le diagnostic de maladie d'Alzheimer à partir de résultats d'analyse du liquide céphalorachidien de patients et de participants sains. Les données d'entraînement proviennent d'une étude de Craig-Schapiro *et al.* (39,40). La prédiction dépend du chemin suivi au sein de l'arbre pour les nouvelles valeurs observées.

2.3.2 Description du mésusage

Au moins deux types de mésusages sont en lien avec l'interprétabilité des modèles d'apprentissage automatique. Le premier concerne la phase de conception d'un modèle, et consiste à ignorer l'importance de l'explicabilité de ce dernier (dans une ou plusieurs de ses acceptions) pour l'application considérée. Le second concerne la phase d'utilisation par les non-spécialistes. Il consiste à accorder un crédit excessif aux prédictions d'un modèle non explicable simplement parce que l'effet de « boîte noire » et la méconnaissance du fonctionnement technique des algorithmes suggèrent des mécanismes de raisonnement analogues à ceux de l'intelligence humaine.

2.3.3 Recommandation pour le chercheur

Lorsqu'on souhaite construire un modèle d'apprentissage automatique, il est recommandé d'évaluer attentivement le niveau d'explicabilité souhaitable à la lumière des enjeux sous-jacents, notamment éthiques et réglementaires, ainsi que du public et du contexte d'utilisation visés. Si les niveaux d'interprétabilité attendus sont élevés, recourir à des classes de modèles tels que régressions linéaires ou logistiques classiques, voire arbres de décision peut être judicieux. Il est sinon possible d'avoir recours à des méthodes d'analyse poussée pour éclaircir le comportement de modèles autrement peu interprétables : graphiques de dépendance partielle, cartes de saillance (41) ou Score-CAM (42) pour les réseaux de neurones, etc. (43). Certaines approches sont utilisables, quelle que soit la classe de modèles, notamment celles reposant sur des approximations linéaires ou par arbres valides localement : on peut citer les approches de type LIME (*Local Interpretable model-agnostic explanations*) (44) ou SHAP (*Shapley additive explanations*) (45) qui sont parmi les plus répandues.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

2.3.4 Recommandation pour le clinicien

Face à des modèles peu explicables (dans un sens à préciser selon le contexte d'utilisation), l'attitude à adopter est celle de la prudence. En effet, le comportement de ces modèles peut être difficile à anticiper, notamment pour des valeurs situées dans la queue de la distribution modélisée. Lorsque le modèle entraîné a été soumis à des analyses d'explicabilité globale, il est conseillé de se faire aider par un statisticien ou scientifique des données pour en interpréter les résultats, au risque d'en tirer des conclusions erronées. Enfin, il ne faut pas perdre de vue l'écart sémantique qui existe entre l'IA et l'intelligence humaine : les modèles d'IA sont des outils statistiques d'optimisation visant à faire des prédictions, sans qu'un raisonnement à proprement parler en soit à l'origine. Il est d'ailleurs possible de manipuler les prédictions de nombreux types de modèles en modifiant subtilement leurs données d'entrée : on peut alors « faire dire » à l'algorithme précisément ce que l'on veut (46).

2.4 Polyvalence et optimalité

2.4.1 Contexte

Il existe aujourd'hui de nombreuses classes de modèles d'apprentissage automatique, utilisées en routine dans une grande variété de domaines. Certaines sont rudimentaires, comme la régression linéaire. D'autres reposent sur des algorithmes plus complexes, et ont permis des gains de performance majeurs pour certaines applications. On peut citer les réseaux de neurones profonds pour le traitement automatique des langues, et les réseaux antagonistes génératifs pour la génération de contenu (notamment d'images). Pour des problèmes de classification ou de régression plus classiques sur données tabulaires (par opposition, par exemple, à des données textuelles ou des images), les algorithmes de *gradient boosting*, comme XGBoost (47), figurent régulièrement parmi les vainqueurs de compétitions internationales (48). Ces réussites pourraient faire espérer l'existence d'un algorithme d'apprentissage automatique optimal pour tous les problèmes d'apprentissage automatique, qui n'attendrait que d'être découvert. Malheureusement, des résultats théoriques forts montrent qu'un tel algorithme ne peut exister. En vertu des théorèmes dits « no free lunch », on peut imaginer, pour tout modèle d'apprentissage automatique, une distribution de données telle que le gain de performance du modèle permis par l'ajustement sur des données de plus grande taille est arbitrairement faible (49). Par ailleurs, comme mentionné dans les sections précédentes, toutes les classes de modèles ne sont pas équivalentes vis-à-vis d'autres caractéristiques souhaitables telles que l'explicabilité ou la robustesse.

2.4.2 Description du mésusage

Les concepteurs ou utilisateurs de modèles peuvent être tentés de privilégier une classe de modèles plutôt que d'autres, uniquement à cause des performances et qualités connues (ou promues) de cette classe pour une catégorie de problèmes et un scénario d'utilisation donnés. Si ceux-ci ne correspondent pas à l'utilisation qui est envisagée pour le modèle à l'étude, alors il s'agit d'un mésusage. Le risque associé est notamment celui d'un manque de performance, les caractéristiques du problème n'ayant pas été suffisamment prises en compte dans ce choix. Ce risque peut concerner d'autres caractéristiques des modèles : il peut s'agir par exemple d'un manque d'explicabilité, pour une application qui pourtant le nécessiterait.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

2.4.3 Recommandation pour le chercheur

La recommandation principale pour les concepteurs de modèles est celle de la polyvalence. En effet, s'il n'existe pas de classe optimale de modèles pour l'ensemble des problèmes d'apprentissage automatique, il en existe pour des catégories de problèmes bien définies (entre autres par des hypothèses sur la loi de probabilité modélisée).
Il s'agit donc de connaître les algorithmes les plus adaptés à chaque situation rencontrée, ce qui peut nécessiter une expertise spécifique allant au-delà de compétences générales en apprentissage automatique. Par exemple, les réseaux de neurones (ainsi que d'autres méthodes comme celles à noyaux) peuvent atteindre un taux de convergence optimal pour des problèmes de régression où la fonction cible est suffisamment régulière, mais seulement sous réserve d'une hyperparamétrisation adéquate (50). Un corollaire est que le chercheur doit connaître les limites de ses compétences techniques en IA (et notamment de son « répertoire » de modèles) afin de solliciter, si nécessaire, un spécialiste du type de problèmes auquel il est confronté.
Par ailleurs, il faut noter que la qualité des prédictions n'est pas la seule caractéristique souhaitable d'un modèle d'apprentissage automatique. À niveau de performance garanti, il vaut mieux, par exemple, privilégier l'apprentissage non profond à l'apprentissage profond. On gagne ainsi, notamment, en « maitrisabilité », opérabilité, transférabilité, et explicabilité. Cette remarque n'implique pas, cependant, qu'un compromis doive systématiquement être trouvé entre ces caractéristiques. Par exemple, il n'est pas établi que les modèles les plus performants pour les types de tâches les plus fréquents soient nécessairement peu explicables (51). Enfin, il faut noter que les déterminants de la fiabilité d'une IA médicale sont avant tout dans la prise en compte de la chaîne complète plutôt que dans le choix d'un algorithme plutôt qu'un autre. Les aspects à prendre en compte comprennent alors la performance, la robustesse, l'explicabilité, la qualité de la base d'apprentissage et la transparence des annotations.

2.4.4 Recommandation pour le clinicien

Une recommandation analogue peut se décliner pour les utilisateurs de modèles. Il s'agit alors de garder une attitude critique face à des propositions techniques dont l'intitulé est séduisant de prime abord. Par exemple, les approches d'apprentissage profond ont gagné en popularité ces dernières années auprès des non-spécialistes, ce qui a pu motiver des tentatives d'application à une vaste gamme de problèmes. L'expérience montre pourtant qu'elles ne sont pas adaptées à tous ces cas d'usage (52). Face à un modèle qualifié de « réseau de neurones profond », deux questions doivent donc être posées. D'une part, il faut savoir si l'algorithme utilisé est, en effet, un algorithme d'apprentissage profond. Mais, surtout, il faut se demander si cette approche est la plus adéquate pour le cas d'usage envisagé. Ces deux questions restent pertinentes, quelle que soit la classe de modèles proposée.
Les outils mentionnés dans les sections précédentes, en plus d'éventuels résultats théoriques d'optimalité, peuvent être utilisés pour y répondre. Notons d'ailleurs que le clinicien peut jouer un rôle essentiel dans la définition de ces cas d'usage, ce qui doit idéalement passer par l'établissement de cahiers des charges, de critères définissant des contextes d'utilisation, de profils d'utilisateurs, mais aussi d'exigences concrètes pour satisfaire les usagers et maîtriser les risques. Pour les cliniciens concernés, ces enjeux opérationnels se déclinent au niveau politique, en vue notamment de la certification

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

des IA pour répondre aux enjeux de souveraineté et d'indépendance technologique. Il s'agit alors, entre autres, d'établir des standards et de contribuer aux bases de données souveraines, ce qui nécessite dans les deux cas d'avoir assimilé les aspects techniques mentionnés ci-dessus.

2.5 Confidentialité

2.5.1 Contexte

De nombreux modèles d'IA pré-entraînés sont aujourd'hui disponibles librement, soit sous la forme de copies pouvant être exécutées localement, soit en tant que services offrant d'interroger le modèle (hébergé sur des serveurs distants) à l'aide de requêtes personnalisées. Dans ce dernier cas, les données fournies par les utilisateurs peuvent elles-mêmes servir à améliorer le modèle. C'est, par exemple, le cas avec ChatGPT (53). Quel que soit le mode de diffusion d'un modèle pré-entraîné, il existe un risque d'atteinte à la confidentialité de ses données d'entraînement par le biais d'attaques par inversion de modèle. Ces attaques font l'objet d'une littérature conséquente depuis quelques années. Elles visent à inférer les caractéristiques des données ayant servi à entraîner un modèle en l'interrogeant de façon répétée et, lorsqu'elles sont connues, en s'aidant de ses propriétés statistiques. Le taux de succès de telles attaques est notamment en lien avec le niveau d'ajustement du modèle : à l'extrême, un modèle complètement surajusté ne peut prédire que les valeurs observées dans ses données d'entraînement, facilitant considérablement l'attaque (54). Quoique de telles attaques puissent paraître anodines de prime abord, elles peuvent avoir des conséquences significatives lorsque les données sont sensibles, notamment en médecine. On peut imaginer, par exemple, qu'une attaque contre un modèle prédisant la séropositivité au VIH permette d'obtenir l'identité des patients séropositifs ayant servi à l'entraîner. Il s'agirait alors d'une infraction grave au droit des données de santé.

2.5.2 Description du mésusage

Le mésusage consiste ici à concevoir des modèles d'IA trop vulnérables à des attaques pouvant porter atteinte à la confidentialité des données d'entraînement. Il se décline du côté utilisateur, lorsque l'interaction avec un modèle fait courir un risque analogue.

2.5.3 Recommandation pour le chercheur

Avant de diffuser un modèle d'IA, il convient donc de s'interroger sur les risques de rupture de confidentialité induits. Des modèles d'analyse spécifiquement conçus à cet effet existent (55). Formellement, on peut considérer que la publication de modèles pré-entraînés s'accompagne d'un niveau de risque intermédiaire entre la publication de statistiques descriptives et celle de microdonnées (c'est-à-dire de données à l'échelle individuelle). Les risques de rupture de confidentialité associés à ces dernières sont particulièrement étudiés dans le cadre des données de haute dimension (où le nombre de variables mesurées dépasse largement le nombre d'individus). Des modèles de confidentialité ont, par exemple, été développés pour les données issues d'études d'association pangénomique, c'est-à-dire à partir des données du génome d'individus par GWAS (56). Ils dérivent classiquement du principe de confidentialité différentielle, qui vise à fournir des garanties formelles de confidentialité dans le cadre d'interactions avec une base de données (57). Enfin, notons qu'un risque analogue existe

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

pour certains modèles génératifs dont des sorties sont publiées, sans que le modèle lui-même ne le soit. Le champ de recherche est alors celui des données synthétiques. Les méthodes d'évaluation du risque de rupture de confidentialité induit par la publication de données synthétiques sont actuellement en plein développement (58).

2.5.4 Recommandation pour le clinicien

La principale vigilance à avoir en tant qu'utilisateur se rapporte au traitement des requêtes émises vers les applications distantes. Lorsque ce traitement implique des réseaux (c'est-à-dire des liaisons d'interaction) ou serveurs ne fournissant pas de garanties de sécurité suffisantes, mieux vaut s'abstenir d'y avoir recours. Il en va de même lorsque les requêtes émises sont réutilisées pour l'entraînement de modèles. Cela est particulièrement vrai lorsque les requêtes impliquent des données sensibles, par exemple de santé.

3 Exemples en neuroradiologie

3.1 Limites de l'IA en neuroradiologie

Depuis plusieurs années, les progrès techniques en termes de vitesse d'acquisition des images et de résolution anatomique ont conduit à une majoration importante du nombre de demandes d'examens d'imagerie en coupe en urgence. Cette augmentation dépasse d'ailleurs nettement celle du nombre d'admissions dans les services d'urgences : à titre d'exemple, une étude israélienne trouvait une hausse respectivement de 52 % des demandes de scanner et de 23 % des visites aux urgences entre 2012 et 2019, soit plus du double (59). Puisqu'une plus grande proportion de patients admis bénéficie d'une imagerie, il en résulte (à incidence constante des pathologies considérées) une augmentation de la proportion d'examens normaux.
Devant ce problème, l'intelligence artificielle pourrait apparaître comme une solution idéale : si sa capacité à raisonner devant des cas pathologiques complexes où de multiples paramètres sont à prendre en compte et plusieurs hypothèses diagnostiques à envisager et à hiérarchiser reste encore à prouver, l'objectif de savoir reconnaître la normalité ou la pathologie simple semble à première vue plus réaliste. Les radiologues pourraient ainsi déléguer cette nouvelle charge de travail conséquente et généralement perçue comme répétitive en se contentant de valider les résultats algorithmiques, et concentrer leur temps vers ces mêmes cas plus difficiles.

Il est donc légitime de s'interroger sur les aptitudes actuelles des modèles d'IA à interpréter ces examens d'imagerie simples afin de s'assurer qu'un tel système n'engendre pas une dégradation de la qualité de la prise en charge. Lorsque l'on compare le taux de réussite de l'IA seule à celui des radiologues, comme l'a fait une étude (parue en 2024) portant sur la détection des fractures de la colonne cervicale en tomodensitométrie (TDM), les résultats sont clairs : elle présentait une sensibilité nettement moins bonne, seulement 71,5 % contre 88,2 % respectivement. Il est toutefois intéressant de noter que, parmi les fractures que l'IA a su repérer, se trouvait la grande majorité (22 sur 26) de celles qui n'avaient pas été vues par les radiologues (60).
Même si cette dernière constatation suggère que l'IA pourrait apporter une approche complémentaire de l'analyse humaine, ses performances globales dans ce cas d'usage semblent encore insuffisantes pour la laisser agir en autonomie.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

L'utilisation de l'IA dans l'interprétation ne peut donc se concevoir à l'heure actuelle qu'en complémentarité à celle du radiologue pour les domaines où les performances de ce dernier ne sont pas exceptionnelles. De tels domaines pourraient, par exemple, inclure les situations diagnostiques complexes rencontrées dans des environnements exposés à des contraintes organisationnelles fortes, comme en réanimation ou dans les services d'urgence, ou encore en opération extérieure. Dans ces situations, les praticiens peuvent être sujets à des biais qui s'expliquent non seulement par ces contraintes, mais aussi par des stéréotypes et le caractère « saillant » de certains signes (c'est-à-dire de signes repérables facilement et spontanément, mais de valeur prédictive potentiellement faible). L'intérêt de l'IA pour améliorer les performances des praticiens dans ces contextes est illustré par plusieurs travaux expérimentaux, dont certains suggèrent aussi un intérêt sur le plan médico-économique (61,62). Au contraire, dans l'exemple (très simple) de la recherche d'hémorragie intracrânienne en TDM, une grande étude prospective de 2021 montrait que les performances des radiologues seuls et assistés par l'IA ne différaient pas significativement, et que cette aide ne raccourcissait pas non plus significativement le délai moyen de complétion des comptes-rendus (63).
S'il convient de préciser que les performances du radiologue seul étaient ici déjà très élevées (« *accuracy* » de 99,5 %) et qu'il semble par conséquent difficile d'obtenir une amélioration, le bénéfice en termes d'efficacité (temps d'interprétation) n'est pas pour autant démontré.

Mais, attention, dans des situations où la performance humaine est moins bonne en raison d'une plus grande part d'incertitude et de subjectivité, cette aide peut parfois se montrer préjudiciable. Ainsi, en se fondant sur la détection des tumeurs pulmonaires en radiographie thoracique, il semble que les radiologues pourraient commettre davantage d'erreurs lorsqu'ils bénéficient de l'assistance d'une IA que seuls. La différence semble d'ailleurs encore plus prononcée lorsque les prédictions du modèle sont sauvegardées dans les dossiers médicaux des patients (64).
D'autres travaux montrent que le niveau de performance de radiologues assistés par l'IA peut s'avérer inférieur, non seulement à celui de radiologues travaillant sans IA, mais également à celui des modèles d'IA utilisés seuls. L'analyse de ce phénomène selon une approche Bayésienne suggère qu'il pourrait résulter d'un biais dans le processus d'inférence mis en œuvre par les radiologues. En considérant les prédictions des modèles d'IA comme indépendantes de leurs propres interprétations, ils aboutiraient, en effet, à des conclusions erronées. Ces biais ne semblent pas concerner les situations où des informations contextuelles sont fournies aux radiologues. Dans ce dernier cas, la performance diagnostique atteinte par les radiologues est améliorée (comme on s'y attendrait) par la prise en compte de ces informations (65). On parle notamment de biais d'automatisation.

Ces expériences montrent qu'au-delà d'une simple addition des performances diagnostiques des deux parties, des facteurs psychologiques plus complexes entrent en jeu dans l'interaction médecin-IA. Il est donc nécessaire que les praticiens acquièrent une plus grande expérience dans l'interprétation assistée par IA afin d'apprendre à estimer correctement ses performances, à les critiquer et à les mettre en contexte, et que cette assistance soit pensée en fonction des biais cognitifs précédemment décrits. La prise en compte de l'incertitude des agents (praticiens et modèles d'IA) jouerait alors un rôle certain. Par exemple, on pourrait imaginer organiser la collaboration

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

radiologue-IA en fonction du niveau d'incertitude de cette dernière (65). Seuls les cas incertains seraient présentés au radiologue, mais en aveugle de la prédiction du modèle d'IA. Notons que ce type d'organisation soulève des questions éthiques et réglementaires, notamment quant à la responsabilité juridique devant être engagée en cas d'erreur diagnostique : doit-elle être celle du radiologue, des concepteurs du système ou autre (66) ? Des collectifs comme Ethik-IA proposent des réponses à ces défis par le concept de « Garantie Humaine », c'est-à-dire la mise en place de points de supervision humaine aux étapes cruciales de la conception et de l'utilisation des modèles d'IA (67).
Un positionnement analogue de l'IA concerne la prédiction de la reproductibilité des diagnostics. Des travaux expérimentaux montrent la faisabilité technique du déploiement de modèles d'IA pour prédire le désaccord entre praticiens, par exemple pour des diagnostics de sévérité émis à partir de rétinophotographies (68). On peut imaginer s'appuyer sur ces modèles pour décider des cas devant faire l'objet d'une discussion collégiale ou d'un second avis. Dans un tel cas d'usage, la démarche diagnostique elle-même n'est alimentée que par un médecin, ou un groupe de médecins, ce qui prévient certains écueils décrits dans cette section.
Malgré les bénéfices qu'une telle aide pourrait apporter, on peut se demander si elle ne risque pas de s'accompagner d'une perte de compétences du radiologue. Elle pourrait, en effet, lui faire perdre l'habitude de raisonner et d'émettre des propositions diagnostiques pertinentes pour toute la diversité de cas radiologiques possibles.

À ces questions concrètes sur les apports de l'IA dans l'interprétation radiologique viennent s'ajouter de nombreuses considérations éthiques (69). Certes, les capacités d'analyse des images par l'IA, plus exhaustive que ce qui est perçu par l'œil humain, pourraient permettre d'identifier des phénomènes non remarqués par le radiologue et en faire un complément précieux. Cependant des dérives peuvent résulter de la prise en compte d'un trop grand nombre d'indicateurs, visibles ou pas. La tendance au surdiagnostic s'en trouverait déjà accélérée. Elle va pourtant déjà croissant depuis les progrès récents de l'imagerie qui conduisent à la découverte d'anomalies parfois subtiles et cliniquement silencieuses, pouvant donner lieu à des traitements et à des investigations inutiles voire délétères. En extrapolant, il est même possible d'imaginer que l'IA devienne un jour capable de repérer précocement certaines maladies avant qu'elles ne soient symptomatiques, ce qui exigerait de disposer d'attitudes cliniques adaptées à ce type de situation (développement de stratégie préventive, par exemple). Il faut aussi penser à des indicateurs intelligibles permettant le développement d'une nouvelle sémiologie infravisible pour des images sous le seuil de détectabilité du radiologue humain, qui serait alors incapable de les critiquer avec, en cas d'erreur de détection ou algorithmiques ou de faux positif, les conséquences graves d'une annonce diagnostique inappropriée qu'aucun autre élément clinique ne pourrait remettre en cause (70).

Enfin, bien que de nombreux outils d'IA soient aujourd'hui validés par les autorités sanitaires, il n'existe que peu d'évaluations scientifiques disponibles. Cela interroge sur la pertinence clinique de leur utilisation en l'absence de démonstration de leur impact sur la prise en charge des patients (71).

De plus, de nombreuses informations non spécifiques aux outils d'IA en neuroradiologie, telles que la validité des bases de données ou les algorithmes utilisés pour leur

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

développement, ne sont souvent pas accessibles aux utilisateurs. Ceci contribue à faire de ces outils des sortes de boîtes noires dont l'utilisation en routine clinique peut être éthiquement discutable d'autant plus qu'elle n'est pas toujours validée selon une méthodologie rigoureuse et explicite (69,72,73)

3.2 Risques liés à la génération d'images en radiologie

L'IA générative est un sujet d'actualité de premier plan depuis plusieurs années, ouvrant de nombreuses perspectives dans de multiples domaines tout autant qu'elle soulève d'importantes questions éthiques. Pour la génération d'images, les progrès récents des algorithmes, aussi appelés réseaux antagonistes génératifs (en anglais : *Generative Adversarial Networks, GAN*), ont permis d'atteindre un haut degré de réalisme. Si ces progrès laissent entrevoir des applications prometteuses dans le domaine médical, ils s'accompagnent également de risques liés au mésusage de ces IA génératives.

Ce mésusage peut être intentionnel. Il consisterait, pour un utilisateur externe malveillant, à altérer le contenu d'examens d'imagerie médicale dans le but d'entraîner des erreurs diagnostiques que le radiologue ou d'autres cliniciens ne seraient en mesure de remettre en cause. Bien que ce risque reste théorique à l'heure actuelle, aucune attaque d'envergure de ce type n'ayant été rapportée, il est bien réel au vu des performances et de l'accessibilité des technologies existantes (74). Il pourrait s'agir d'altérations d'images ayant pour objet d'induire en erreur un observateur humain par l'ajout ou le retrait de signaux sains ou de lésions visibles ou détectées uniquement par l'IA telles que des tumeurs, par exemple. Ces risques sont d'autant plus sévères que la technologie permettant d'effectuer ces manipulations complexes d'images est devenue accessible à travers les smartphones de dernière génération, réduisant considérablement les compétences techniques nécessaires ainsi que la durée du processus (75).

Ce procédé de transformation intentionnelle des images peut rester très efficace même lorsque les radiologues sont avertis de l'attaque et de la possibilité de fausses images (76), ce qui appelle un renforcement des moyens de cybersécurité pour des applications médicales. Les mesures de protection par IA font notamment l'objet de recherches croissantes : il s'agit de l'utilisation de modèles conçus pour détecter des artéfacts subtils présents sur ces images générées artificiellement (77) dont la performance en vie réelle doit idéalement être monitorée en continu (74).

Comme nous l'avons écrit plus haut, l'IA générative nécessite un entraînement à partir d'une base de données dont la taille détermine les performances qui seront atteintes. Le devenir de ces données soulève d'autres préoccupations : si ces données sensibles sont imparfaitement anonymisées, alors le modèle peut les apprendre et donc potentiellement les reproduire, à l'instar de texte flouté accompagnant les images et contenant des informations personnelles que certaines IA parviennent à déchiffrer. Même en limitant au maximum ces risques, il reste que les images sont en elles-mêmes de nature sensible, car elles peuvent permettre d'effectuer des reconstructions 3D à partir d'un examen encéphalique et de reconstituer le visage d'un patient, par exemple.

Un autre versant du champ d'application de l'IA générative en radiologie est celui de la reconstruction des images d'examens en coupe. En IRM notamment, elle peut permettre de diminuer les temps d'acquisition en « prédisant » l'apparence d'images de

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

haute qualité à partir de celles acquises plus rapidement avec une qualité volontairement diminuée. Plusieurs études ont, en effet, montré que cette approche permettait de réduire notablement la durée des examens tout en conservant des performances diagnostiques superposables à celles des acquisitions standards (78). Néanmoins, celles-ci ont été le plus souvent conduites sur des échantillons de volontaires sains et évaluées sur un type d'examen unique pour lequel les algorithmes de reconstruction étaient optimisés. Les conséquences imprévues apportées par la variété des pathologies rencontrées et les conditions plus difficiles d'acquisition en pratique clinique réelle restent donc une inquiétude légitime et doivent encore être évaluées (79).

Ainsi, malgré les solutions théoriquement apportées par l'IA pour résoudre le problème de l'augmentation de la charge de travail des radiologues, elle pourrait également contribuer à l'aggraver si ces questions ne sont pas correctement traitées (80).

Conclusion

Si l'IA est une formidable alliée de l'intelligence humaine pour les tâches répétitives, la manipulation d'un nombre presque illimité de paramètres et de leurs combinaisons permettant ainsi de disposer de modèles dans des problèmes aux dimensions dépassant l'entendement humain, elle reste fortement tributaire des faiblesses intrinsèques à sa constitution ou de failles de son utilisation dans le domaine médical. Il est pour cela important de comprendre ces faiblesses et de les prendre en compte. Le praticien en général, mais qui plus est militaire, doit aussi être en mesure de comprendre et de connaître les limites de son application pour éviter à tout prix de l'utiliser dans des champs qui ne lui sont pas adaptés. Ces mésusages peuvent, en effet, conduire à des résultats erronés, du fait de biais complexes issus de techniques opaques pour un public non initié, revêtant l'apparence de faits solidement défendus.

Remerciements

Les auteurs remercient chaleureusement Augustin Lecler, Mickaël Ohana
et Félix Renard pour les discussions fertiles qui ont parachevé le caractère collaboratif de ce travail.

Remerciements au linguiste relecteur : Pierre-Alexandre Clerc.

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*,
Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

Références bibliographiques

1. Ahmad Z, Rahim S, Zubair M, Abdul-Ghafar J. Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology : present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. Diagnostic Pathology. 2021 ;16(1) :24. doi : 10.1186/s13000-021-01086-4
2. Biswas S. ChatGPT and the future of medical writing. Radiology. 2023 ;307(2) :e223312. doi : 10.1148/radiol.223312
3. Olivier M, Asmis R, Hawkins GA, Howard TD, Cox LA. The need for multi-omics biomarker signatures in precision medicine. International Journal of Molecular Sciences. 2019 ;20(19) :4781. doi : 10.3390/ijms20194781
4. Xian X. Unveiling the future : how can GAI transform mental health care ? Journal of Medical Humanity and Media. 2024 ;2(1) :76-93. doi : 10.1007/s44280-024-00016-x
5. Lee JD, See KA. Trust in automation : designing for appropriate reliance. Human Factors. 2004 ;46(1) : 50-80. doi : 10.1518/hfes.46.1.50_30392
6. Lyell D, Magrabi F, Raban MZ, Pont LG, Baysari MT, Day RO et al. Automation bias in electronic prescribing. BMC Medical Informatics and Decision Making. 2017 ;17(1) :28. doi : 10.1186/s12911-017-0425-5
7. Dean T, Allen J, Aloimonos Y. Artificial intelligence : theory and practice. Redwood City (US) : Benjamin-Cummings Publishing ; 1995.
8. Jackson PC. An introduction to artificial intelligence. New York (US) : Dover Publications ; 1985.
9. Jackson P. Introduction to expert systems. Reading (UK) : Addison-Wesley ; 1986.
10. Shao J. Mathematical statistics. New York (US) : Springer ; 2008. doi : 10.1007/978-0-387-24970-4
11. Azencott CA. Introduction au machine learning. 2e éd. Paris (FR) : Dunod ; 2022.
12. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning : data mining, inference, and prediction. New York (US) : Springer ;2017. doi : 10.1007/978-0-387-84858-7
13. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015 ;521(7553) :452-9. doi : 10.1038/nature14541
14. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 ;521(7553) :436-44. doi : 10.1038/nature14539
15. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. Business & Information Systems Engineering. 2024 ;66(1) :111-26. doi : 10.1007/s12599-023-00834-7
16. Fetzer JH. What is artificial intelligence ? Dans : Fetzer JH, éditeur. Artificial intelligence : its scope and limits. Dordrecht (NL) : Springer ; 1990. p.3-27. doi : 10.1007/978-94-009-1900-6_1
17. French RM. The Turing test : the first 50 years. Trends in Cognitive Sciences. 2000 ;4(3) :115-22. doi : 10.1016/s1364-6613(00)01453-4
18. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J et al. A survey of large language models in medicine : progress, application, and challenge. arXiv. 2024. Disponible : http://arxiv.org/abs/2311.05112
19. Chen Z, Romanou A, Bonnet A, Hernández-Cano A, Alkhamissi B, Matoba K et al. MEDITRON : open medical foundation models adapted for clinical practice. Research Square ; 2024. Disponible : https://www.researchsquare.com/article/rs-4139743/latest
20. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M *et al.* Toward expert-level medical question answering with large language models. Nature Medicine. 2025 ;31(1) :1-8. doi : 10.1038/s41591-024-03301-3
21. Shortliffe E. Computer-based medical consultations : MYCIN. New York(US) :Elsevier ;2012. doi : 10.1016/C2013-0-10095-0
22. Lehmann EL, Romano JP. The general decision problem. Dans : Lehmann EL, Romano JP. Testing statistical hypotheses. Cham (CH) : Springer ; 2022. p.3-28. doi : 10.1007/978-3-030-70578-7_1
23. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys. 2022 ;54(6) :1-35. doi : 10.1145/3457607
24. Brodley CE, Friedl MA. Identifying mislabeled training data. Journal of Artificial Intelligence Research. 1999 ;11 :131-67. doi : 10.1613/jair.606
25. Natarajan N, Dhillon IS, Ravikumar PK, Tewari A. Learning with noisy labels. Advances in Neural Information Processing Systems. 2013 ;26 :1196-204. Disponible : https://proceedings.neurips.cc/paper/5073-learning-with-noisy-labels
26. Pessach D, Shmueli E. A review on fairness in machine learning. ACM Computing Surveys. 2023 ;55(3) : 1-44. doi : 10.1145/3510415

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

27. Xu H, Caramanis C, Mannor S. Robustness and regularization of support vector machines. Journal of Machine Learning Research. 2009 ;10 :1485-510. Disponible : https://www.jmlr.org/papers/volume10/xu09b/xu09b.pdf

28. Jaipuria N, Zhang X, Bhasin R, Arafa M, Chakravarty P, Shrivastava S et al. Deflating dataset bias using synthetic data augmentation. Dans : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops ; 2020. New York (US) : IEEE ; 2020. p.772-3. Disponible : http://openaccess.thecvf.com/content_CVPRW_2020/html/w45/Jaipuria_Deflating_Dataset_Bias_Using_Synthetic_Data_Augmentation_CVPRW_2020_paper.html

29. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big Data. 2016 ;3(1) :9. doi : 10.1186/s40537-016-0043-6

30. Falissard B. Comprendre et utiliser les statistiques en médecine. Paris(FR) : Elsevier Masson ; 2023.

31. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna : a next-generation hyperparameter optimization framework. Dans : Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining ; 2019 ; USA. New York(US) :ACM ;2019. p.2623-31. doi : 10.1145/3292500.3330701

32. Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt : a python library for model selection and hyperparameter optimization. Computational Science & Discovery. 2015 ;8(1) :014008. doi : 10.1088/1749-4699/8/1/014008

33. Ying X. An overview of overfitting and its solutions. Journal of Physics : Conference Series. 2019 ;1168 :022022. doi : 10.1088/1742-6596/1168/2/022022

34. Jiang Y, Neyshabur B, Mobahi H, Krishnan D, Bengio S. Fantastic generalization measures and where to find them. arXiv. 2019. Disponible : http://arxiv.org/abs/1912.02178

35. Burkart N, Huber MF. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research. 2021 ;70 :245-317. doi : 10.1613/jair.1.12228

36. Kundu S. AI in medicine must be explainable. Nature Medicine. 2021 ;27(8) :1328. doi : 10.1038/s41591-021-01461-z

37. Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence. 2021 ;5(5) :726-42. doi : 10.1109/TETCI.2021.3070808

38. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R *et al.* DeepSeek-R1 : incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. 2025. Disponible : http://arxiv.org/abs/2501.12948

39. Craig-Schapiro R, Kuhn M, Xiong C, Pickering EH, Liu J, Misko TP *et al.* Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and prognosis. PLoS ONE. 2011 ;6(4) :e18850. doi : 10.1371/journal.pone.0018850

40. Kuhn M, Johnson K. Applied predictive modeling. New York(US) : Springer ;2013. doi : 10.1007/978-1-4614-6849-3

41. Hong S, You T, Kwak S, Han B. Online tracking by learning discriminative saliency map with convolutional neural network. Dans : International Conference on Machine Learning ; 2015 ; France. New York (US) : PMLR ; 2015. p. 597-606. Disponible : https://proceedings.mlr.press/v37/hong15.html

42. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S et al. Score-CAM : score-weighted visual explanations for convolutional neural networks. Dans : Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops ; 2020 ; Virtual. New York(US) : IEEE ; 2020. p.24-5. Disponible : http://openaccess.thecvf.com/content_CVPRW_2020/html/w1/Wang_Score-CAM_Score-Weighted_Visual_Explanations_for_Convolutional_Neural_Networks_CVPRW_2020_paper.html

43. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability : a survey on methods and metrics. Electronics. 2019 ;8(8) :832. doi : 10.3390/electronics8080832

44. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you ?" : explaining the predictions of any classifier. Dans : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ; 2016 ; USA.New York(US) :ACM ;2016. p. 1135-44. doi : 10.1145/2939672.2939778

45. Rodríguez-Pérez R, Bajorath J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. Journal of Medicinal Chemistry. 2020 ;63(16) :8761-77. doi : 10.1021/acs.jmedchem.9b01101

46. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 2019 ;363(6433) :1287-9. doi : 10.1126/science.aaw4399

47. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H et al. Xgboost : extreme gradient boosting. R Package Version 0.4-2. 2015 ;1(4) :1-4. Disponible : https://cran.r-project.org/web/packages/xgboost/index.html

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

48. Shwartz-Ziv R, Armon A. Tabular data : deep learning is not all you need. Information Fusion. 2022 ;81 :84-90. doi : 10.1016/j.inffus.2021.11.011

49. Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition. New York (US) : Springer ; 2013. doi : 10.1007/978-1-4612-0711-5

50. Stone CJ. Optimal global rates of convergence for nonparametric regression. The Annals of Statistics. 1982 ;10(4) :1040-53. doi : 10.1214/aos/1176345969

51. Rudin C, Radin J. Why are we using black box models in AI when we don't need to ? A lesson from an explainable AI competition. Harvard Data Science Review. 2019 ;1(2) :1-9. doi : 10.1162/99608f92.e26845b2

52. Ahmed M, Najmul Islam AKM. Deep learning : hope or hype. Annals of Data Science. 2020 ;7(3) :427-32. doi : 10.1007/s40745-020-00274-0

53. Wu X, Duan R, Ni J. Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information Intelligence. 2024 ;2(2) :102-15. doi : 10.1016/j.jinfint.2023.11.001

54. Dibbo SV. SoK : model inversion attack landscape : taxonomy, challenges, and future roadmap. Dans : 2023 IEEE 36th Computer Security Foundations Symposium (CSF) ; 2023 ; Virtual. New York (US) :IEEE ; 2023.p.439-56. doi : 10.1109/CSF57540.2023.00010

55. Qiu Y, Yu H, Fang H, Yu W, Chen B, Wang X et al. MIBench : a comprehensive benchmark for model inversion attack and defense. arXiv. 2024. Disponible : http://arxiv.org/abs/2410.05159

56. Fienberg SE, Slavkovic A, Uhler C. Privacy preserving GWAS data sharing. Dans : 2011 IEEE 11th International Conference on Data Mining Workshops ; 2011 ; Canada. New York (US) : IEEE ; 2011.p.628-35. doi : 10.1109/ICDMW.2011.77

57. Dwork C. Differential privacy. Dans : Bugliesi M, Preneel B, Sassone V, Wegener I, éditeurs. Automata, languages and programming. Berlin (DE) : Springer ; 2006. p.1-12. doi : 10.1007/11787006_1

58. Osorio-Marulanda PA, Epelde G, Hernandez M, Isasa I, Reyes NM, Iraola AB. Privacy mechanisms and evaluation metrics for synthetic data generation : a systematic review. IEEE Access. 2024 ;12 :88048-88074. doi : 10.1109/ACCESS.2024.3414368

59. Lantsman CD, Barash Y, Klang E, Guranda L, Konen E, Tau N. Trend in radiologist workload compared to number of admissions in the emergency department. European Journal of Radiology. 2022 ;149 :110195. doi : 10.1016/j.ejrad.2022.110195

60. van den Wittenboer GJ, van der Kolk BY, Nijholt IM, Langius-Wiffen E, van Dijk RA, van Hasselt BA et al. Diagnostic accuracy of an artificial intelligence algorithm versus radiologists for fracture detection on cervical spine CT. European Radiology. 2024 ;34(8) :5057-66. doi : 10.1007/s00330-024-10876-8

61. King AJ, Tang L, Davis BS, Preum SM, Bukowski LA, Zimmerman J et al. Machine learning-based prediction of low-value care for hospitalized patients. Intelligence-Based Medicine. 2023 ;8 :100115. doi : 10.1016/j.ibmed.2023.100115

62. Mullainathan S, Obermeyer Z. Diagnosing physician error : a machine learning approach to low-value health care. The Quarterly Journal of Economics. 2022 ;137(2) :679-727. doi : 10.1093/qje/qjab048

63. Savage CH, Tanwar M, Elkassem AA, Sturdivant A, Hamki O, Sotoudeh H et al. Prospective evaluation of artificial intelligence triage of intracranial hemorrhage on noncontrast head CT examinations. American Journal of Roentgenology. 2024 ;223(5). doi : 10.2214/AJR.24.31639

64. Bernstein MH, Atalay MK, Dibble EH, Maxwell AWP, Karam AR, Agarwal S et al. Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it ? A multi-reader pilot study of lung cancer detection with chest radiography. European Radiology. 2023 ;33(11) : 8263-9. doi : 10.1007/s00330-023-09747-0

65. Agarwal N, Moehring A, Rajpurkar P, Salz T. Combining human expertise with artificial intelligence : experimental evidence from radiology. Cambridge (UK) : National Bureau of Economic Research ; 2023. Disponible : https://www.nber.org/papers/w31422

66. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms : a systematic review. Frontiers in Medicine. 2023 ;10 :1305756. doi : 10.3389/fmed.2023.1305756

67. Gruson D. Artificial intelligence in health care : la machine, le médecin et moi or the world of SARRA ? L'Information Psychiatrique. 2020 ;96(5) :323-30. doi : 10.1684/ipe.2020.2163

68. Raghu M, Blumer K, Sayres R, Obermeyer Z, Kleinberg B, Mullainathan S et al. Direct uncertainty prediction for medical second opinions. Dans : International Conference on Machine Learning ; 2019 ; USA. New York (US) : PMLR ; 2019. p.5281-90. Disponible : https://proceedings.mlr.press/v97/raghu19a

69. Khosravi P, Schweitzer M. Artificial intelligence in neuroradiology : a scoping review of some ethical challenges. Frontiers in Radiology. 2023 ;3 :1149461. doi : 10.3389/fradi.2023.1149461

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816

70. Can AI. Who will be the third umpire : AI or radiologists ? The National Medical Journal of India. 2024 ;37(3) :121-2. Disponible : https://www.nmji.in/content/141/2024/37/3/pdf/NMJI-37-121.pdf
71. Olthof AW, van Ooijen PMA, Rezazade Mehrizi MH. Promises of artificial intelligence in neuroradiology : a systematic technographic review. Neuroradiology. 2020 ;62(10) :1265-78. doi : 10.1007/s00234-020-02424-3
72. Haller S, Van Cauter S, Federau C, Hedderich DM, Edjlali M. The R-AI-DIOLOGY checklist : a practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology. Neuroradiology. 2022 ;64(5) :851-64. doi : 10.1007/s00234-021-02854-7
73. Lui YW, Chang PD, Zaharchuk G, Barboriak DP, Flanders AE, Wintermark M et al. Artificial intelligence in neuroradiology : current status and future directions. American Journal of Neuroradiology. 2020 ;41(8) :E52-E59. doi : 10.3174/ajnr.A6680
74. Sorin V, Soffer S, Glicksberg BS, Barash Y, Konen E, Klang E. Adversarial attacks in radiology : a systematic review. European Journal of Radiology. 2023 ;166 :111085. doi : 10.1016/j.ejrad.2023.111085
75. Duron L, Soyer P, Lecler A. Generative AI smartphones : from entertainment to potentially serious risks in radiology. Diagnostic and Interventional Imaging. 2024 ;105(10) :369-70. doi : 10.1016/j.diii.2024.06.002
76. Mirsky Y, Mahler T, Shelef I, Elovici Y. CT-GAN : malicious tampering of 3D medical imagery using deep learning. Dans : 28th USENIX Security Symposium (USENIX Security 19) ; 2019 ; USA. Berkeley (US) : USENIX Association ; 2019. p.461-78. Disponible : https://www.usenix.org/conference/usenixsecurity19/presentation/mirsky
77. Chu LC, Anandkumar A, Shin HC, Fishman EK. The potential dangers of artificial intelligence for radiology and radiologists. Journal of the American College of Radiology. 2020 ;17(10) :1309-11. doi : 10.1016/j.jacr.2020.07.010
78. Brady SL. Implementation of AI image reconstruction in CT : how is it validated and what dose reductions can be achieved. The British Journal of Radiology. 2023 ;96(1150) :20220915. doi : 10.1259/bjr.20220915
79. Zhang Z, Seeram E. The use of artificial intelligence in computed tomography image reconstruction : a literature review. Journal of Medical Imaging and Radiation Sciences. 2020 ;51(4) :671-7. doi : 10.1016/j.jmir.2020.09.001
80. Harvey H, Topol EJ. More than meets the AI : refining image acquisition and resolution. The Lancet. 2020 ;396(10261) :1479. doi : 10.1016/S0140-6736(20)32209-6

Published in : SERVICE DE SANTÉ DES ARMÉES, - (dir.), *Médecine et Armées Vol.51 No.2. Dossier Numérique en santé*, Editions des archives contemporaines, ISBN : 9782813005816, doi : 10.17184/eac.9782813005816