

REVUE SYSTÉMATIQUE POUR L'UERB25

année universitaire 2025-2026

problématique : Comment les caractéristiques des jeux de données conditionnent-elles les performances des systèmes d'intelligence artificielle en radiographie et leur généralisation en pratique clinique ?

Introduction

Parmi les innovations technologiques ayant marqué la dernière décennie, l'intelligence artificielle (IA) occupe une place prépondérante en raison de son développement rapide et de son adoption dans de nombreux secteurs d'activité. L'IA se définit comme un ensemble de méthodes et de techniques informatiques permettant à un système de réaliser des tâches telles que l'apprentissage, le raisonnement, la reconnaissance de formes, l'analyse, la compréhension et l'aide à la décision. Aujourd'hui, l'IA est présente dans de nombreux secteurs de la société et transforme progressivement les pratiques professionnelles ainsi que les processus décisionnels.

Schématiquement, un système d'IA peut être comparé à une usine qui transforme des données d'entrée en informations ou en prédictions grâce à un processus d'apprentissage. Les performances d'un système d'IA dépendent donc directement des données utilisées pour son entraînement. En d'autres termes, un modèle ne peut apprendre qu'à partir des informations qui lui sont fournies, ce qui confère aux données d'entrée une importance toute particulière si l'on souhaite obtenir des résultats de qualité.

Pour une tâche donnée, ces informations sont généralement regroupées au sein d'un jeu de données (*dataset*), c'est-à-dire un ensemble structuré d'exemples utilisé pour entraîner, valider et tester un modèle d'IA. Chaque réponse produite par l'IA est donc fonction du ou des jeux de données auxquels elle a eu accès. Ce principe est souvent résumé par l'expression « *Garbage In, Garbage Out* », selon laquelle la qualité des résultats produits par un système d'IA dépend directement de la qualité des données utilisées pour son entraînement. Ainsi, les caractéristiques de ces jeux de données (taille, qualité, diversité, représentativité) conditionnent directement les performances, la robustesse et la capacité de généralisation des modèles développés.

En radiologie, l'IA occupe une place croissante, notamment en raison de l'augmentation continue du volume d'examens d'imagerie, de la pénurie de radiologues et de sa capacité à analyser rapidement de grandes quantités d'images médicales. Les performances des systèmes d'IA sont susceptibles d'influencer l'interprétation des examens radiologiques, l'établissement du diagnostic, l'indication d'examens complémentaires et, par conséquent, la prise en charge des patients. Or, les caractéristiques des différentes populations varient considérablement d'une région à l'autre, d'un âge à l'autre ou encore d'un contexte clinique à

l'autre. Un modèle entraîné sur une population donnée ne conserve donc pas nécessairement les mêmes performances lorsqu'il est appliqué à une population différente. Cette capacité, appelée généralisation, constitue aujourd'hui l'un des principaux défis de l'IA en radiologie. Dans ce contexte, il apparaît pertinent d'examiner dans quelle mesure les caractéristiques des jeux de données influencent les performances des systèmes d'intelligence artificielle en radiographie ainsi que leur capacité à généraliser leurs résultats en pratique clinique.

Pour répondre à cette problématique, nous présenterons dans un premier temps les caractéristiques d'un jeu de données permettant le développement d'une IA performante en radiologie. Nous nous intéresserons ensuite aux principales limites actuelles des jeux de données et à leurs conséquences sur les performances des systèmes d'IA. Enfin, nous étudierons les différentes perspectives d'amélioration permettant de développer des modèles plus robustes et davantage généralisables en pratique clinique.

I. Les caractéristiques d'un jeu de données permettant le développement d'une IA radiologique performante

1.1. Les jeux de données au cœur de l'apprentissage des systèmes d'IA

Historiquement, deux grandes approches ont été utilisées pour développer des modèles d'intelligence artificielle destinés à l'analyse des radiographies.

La première est la méthode dite traditionnelle, dans laquelle un expert définit manuellement les caractéristiques à extraire des radiographies (*feature engineering*). L'IA apprend ensuite à partir de ces caractéristiques. Dans cette approche, l'algorithme utilise uniquement des critères déjà connus, puisqu'ils sont définis par l'expert. Il se retrouve donc limité aux caractéristiques analytiques préalablement sélectionnées. En revanche, cette méthode ne nécessite pas de jeux de données très volumineux, puisque l'on connaît précisément les informations analysées et les critères conduisant au résultat.

La seconde approche repose sur les algorithmes d'apprentissage profond (*deep learning*), dans lesquels aucune caractéristique n'est définie au préalable. L'IA ne reçoit aucune consigne concernant les éléments à analyser : elle extrait elle-même les informations contenues dans les images et identifie de façon autonome les caractéristiques les plus pertinentes pour l'analyse radiographique. Cette méthode offre une capacité d'analyse nettement supérieure à celle des approches traditionnelles et permet de détecter des informations beaucoup plus fines, parfois imperceptibles à l'œil humain. En contrepartie, elle nécessite un volume très important de données afin d'obtenir des performances satisfaisantes.

Comme expliqué précédemment, les méthodes d'apprentissage profond présentent des capacités nettement supérieures aux approches traditionnelles. C'est précisément dans

cette optique que la qualité des jeux de données est devenue un élément central du développement de l'intelligence artificielle. Autrefois, avec les modèles traditionnels, l'essentiel du travail consistait à guider l'IA dans son apprentissage. Aujourd'hui, si l'on souhaite développer un modèle performant pour l'analyse radiographique, il est indispensable de lui fournir des jeux de données de la meilleure qualité possible. En d'autres termes, avec le *deep learning*, les données ne constituent plus un simple facteur d'amélioration des performances : elles deviennent une condition essentielle de l'apprentissage, de la généralisation des connaissances et de l'utilisation clinique des modèles. C'est pourquoi de nouveaux standards de qualité des données ont progressivement émergé. Ces nouvelles exigences ont conduit au développement de recommandations méthodologiques visant à améliorer la qualité des données utilisées pour entraîner les modèles de *deep learning*.

1.2. Les caractéristiques d'un jeu de données de qualité

Précédemment, nous avons traité de l'importance d'un jeu de données de qualité. Intéressons-nous maintenant aux caractéristiques nécessaires à l'élaboration de jeux de données destinés à l'analyse de radiographies. Pour appuyer notre argumentaire, nous pouvons établir une analogie entre un jeu de données et un échantillon d'étude de cohorte : de la même manière qu'un échantillon doit être représentatif de la population étudiée, un jeu de données doit représenter la diversité de la population cible afin de conserver ses performances en pratique clinique.

Ainsi, la taille du jeu de données, c'est-à-dire le nombre de radiographies disponibles, joue un rôle important, puisqu'un grand jeu de données a statistiquement davantage de chances de contenir une plus grande variété de cas cliniques. Cette diversité permet à l'IA d'améliorer sa robustesse diagnostique. Il est toutefois important de préciser qu'un jeu de données volumineux ne s'affranchit pas pour autant des biais potentiels. Ceux-ci peuvent néanmoins être limités par une sélection plus représentative des données, intégrant notamment une diversité ethnique, générationnelle, géographique, sexuelle ainsi que des profils de comorbidités variés.

Un jeu de données de qualité se veut également multicentrique. Le multicentrisme permet de limiter l'influence des protocoles propres à chaque établissement, des différences de populations prises en charge (secteur public ou privé) ainsi que de la variabilité des appareils de radiologie utilisés pour réaliser les clichés.

Les annotations associées aux radiographies constituent une autre composante tout aussi essentielle que la taille du jeu de données. Elles sont d'abord nécessaires pour guider l'apprentissage de l'IA en lui transmettant les informations issues de l'interprétation réalisée par le radiologue. Il est ensuite indispensable que ces annotations soient de grande qualité afin de ne pas induire le modèle en erreur. Cette exigence fait directement écho au principe de *Garbage In, Garbage Out* : des annotations erronées empêchent l'IA de construire un raisonnement fiable et compromettent ainsi l'analyse des radiographies. C'est pourquoi une vérification des annotations est généralement réalisée afin d'en garantir la qualité. Dans cette même logique, il est recommandé de faire intervenir plusieurs radiologues pour la

création et la validation des annotations, afin de limiter les biais pouvant être introduits par un seul observateur.

Enfin, une dernière caractéristique indispensable d'un jeu de données de qualité est la prise en compte de l'incertitude diagnostique. En radiographie, comme plus largement en imagerie médicale, il est fréquent qu'un diagnostic soit formulé avec un certain degré d'incertitude. Intégrer cette information permet à l'IA de ne pas considérer chaque annotation comme une vérité absolue et de hiérarchiser les différentes hypothèses diagnostiques. Cette approche contribue à limiter l'impact d'éventuelles erreurs d'annotation et améliore ainsi les performances du modèle. Elle permet également d'inclure un plus grand nombre d'examen dans le jeu de données, contrairement à une approche reposant uniquement sur des diagnostics parfaitement certains.

La qualité d'un jeu de données constitue donc un facteur déterminant des performances des systèmes d'intelligence artificielle. Toutefois, elle ne suffit pas, à elle seule, à garantir leur efficacité en pratique clinique. Il est donc nécessaire d'examiner les autres conditions indispensables au développement d'une IA robuste.

1.3. Les conditions nécessaires à la généralisation des performances en pratique clinique

Nous avons expliqué, dans la partie précédente, l'importance de la qualité des jeux de données utilisés pour entraîner l'IA. Bien que cet aspect soit fondamental, il ne suffit pas à lui seul à construire un modèle utilisable en pratique clinique. Une étape essentielle avant tout déploiement consiste à évaluer objectivement le modèle dans des conditions se rapprochant le plus possible de son utilisation réelle : populations différentes, établissements distincts, protocoles d'acquisition variés ou encore équipements d'imagerie différents.

Cette capacité, appelée généralisation, constitue aujourd'hui l'un des principaux défis du développement des modèles d'IA en radiologie. En effet, un modèle de *deep learning* peut obtenir d'excellents résultats durant sa phase de développement, puis voir ses performances diminuer lorsqu'il est confronté à des conditions réelles d'utilisation. En pratique clinique, il sera inévitablement exposé à des données qu'il n'a jamais rencontrées auparavant. Il est impossible de contrôler l'ensemble des caractéristiques des patients ou des examens qu'il analysera : différences démographiques, prévalence des pathologies, protocoles d'acquisition des images ou encore constructeurs des appareils de radiologie.

À défaut de pouvoir contrôler cette variabilité, il est possible d'évaluer le modèle sur des données totalement indépendantes de celles ayant servi à son entraînement afin d'apprécier sa capacité de généralisation. Cette étape, appelée validation externe, permet de tester l'IA en dehors de son environnement d'apprentissage et d'évaluer ses performances dans de nouvelles conditions.

Au-delà de cette validation externe, il est indispensable d'accorder une attention particulière à la méthodologie d'évaluation des modèles. Lorsqu'on cherche à mesurer les performances d'un algorithme, il est essentiel de décrire avec précision les expériences réalisées, leur

contexte, les jeux de données utilisés, les critères d'évaluation, le niveau d'incertitude ainsi que les éventuelles répétitions des expérimentations. Cette transparence méthodologique est indispensable pour comparer les performances des différents modèles de manière fiable. Dans cette optique, les recommandations CONSORT-AI et SPIRIT-AI proposent un cadre méthodologique destiné à améliorer la qualité des études cliniques portant sur les systèmes d'intelligence artificielle.

Les difficultés rencontrées pour généraliser les performances des modèles ont conduit à un changement de paradigme dans le domaine de l'intelligence artificielle. Alors que l'amélioration des architectures algorithmiques constituait historiquement le principal levier de progrès, la littérature récente considère désormais que les gains de performance reposent avant tout sur l'amélioration de la qualité des jeux de données. Cette approche, appelée data-centric AI, souligne qu'une meilleure qualité des données, des annotations plus fiables et une représentativité accrue des populations constituent aujourd'hui les principaux leviers d'amélioration des systèmes d'intelligence artificielle.

Finalement, un système d'analyse radiographique par IA ne repose pas sur une caractéristique unique, mais sur la combinaison de plusieurs éléments complémentaires : des jeux de données de qualité, une validation multicentrique, une évaluation clinique rigoureuse ainsi qu'une transparence méthodologique garantissant la reproductibilité des résultats.

II. Les limites actuelles des jeux de données et leurs conséquences sur les performances des systèmes d'IA

Bien que les principales étapes nécessaires à la conception d'une IA performante pour l'analyse de radiographies soient aujourd'hui bien établies, le passage de la théorie à la pratique reste complexe en raison des contraintes réelles et du niveau d'exigence requis pour une utilisation en contexte clinique.

2.1. Une généralisation encore insuffisante

Comme évoqué précédemment, la généralisation constitue un critère essentiel de qualité pour une IA fiable et performante. Cependant, elle représente également une limite largement décrite dans la littérature scientifique.

Une revue systématique menée par Kim et al. rapporte que l'évaluation de modèles d'IA sur des jeux de données externes entraîne une baisse significative des performances cliniques. Elle met en évidence une diminution des performances dans 81 % des modèles étudiés, supérieure à 5 % dans la moitié des cas et supérieure à 10 % dans un quart des cas.

Cette dégradation des performances s'explique par la forte variabilité des données rencontrées en pratique clinique, qui diffèrent souvent de celles utilisées lors de l'entraînement des modèles. Ce phénomène, appelé *domain shift*, regroupe notamment les différences de populations, de prévalence des pathologies, de protocoles d'acquisition, de constructeurs d'appareils d'imagerie, ainsi que de qualité des images.

L'IA apprend des relations statistiques à partir des jeux de données d'entraînement et suppose implicitement que ces relations restent valables lors de la phase d'application clinique, alors même qu'elles ne sont pas toujours représentatives de la population cible. Dans ce contexte, les algorithmes peinent à s'adapter, ce qui entraîne une diminution de leurs performances.

Ce manque de généralisation peut être aggravé par le surapprentissage (*overfitting*), au cours duquel le modèle apprend trop spécifiquement les caractéristiques du jeu de données d'entraînement, au détriment de sa capacité à s'adapter à de nouvelles données.

Ainsi, ce déficit de généralisation constitue un obstacle majeur à l'utilisation des modèles de deep learning en radiologie, car il limite leur déploiement généralisé. Pour maintenir des performances acceptables dans différents environnements cliniques, il devient souvent nécessaire d'adapter ou de recalibrer les modèles à chaque nouveau contexte, ce qui revient en pratique à les réentraîner, mobilisant ainsi des ressources humaines et matérielles importantes.

À défaut de cette adaptation, on observe une augmentation des faux positifs et des faux négatifs, réduisant l'efficacité clinique des systèmes. L'objectif initial de ces algorithmes — gain de temps, amélioration de la performance diagnostique et optimisation de la prise en charge des patients — se trouve alors compromis.

2.2. Les biais liés aux jeux de données

La généralisation n'est pas le seul facteur limitant le développement des algorithmes appliqués à l'analyse radiologique. Les biais constituent également une problématique majeure dans la conception de modèles robustes et fiables en population générale.

Les biais correspondent à des erreurs systématiques introduites lors de la conception, de la collecte, de l'annotation ou de l'utilisation des données. Contrairement aux erreurs aléatoires, ils affectent durablement le comportement du modèle et entraînent une dégradation systématique des performances.

Ils sont particulièrement difficiles à identifier, car ils peuvent apparaître à chaque étape du développement d'un système d'IA. De plus, ils peuvent concerner de multiples aspects : biais de sélection, biais démographiques, biais liés aux appareils d'imagerie, biais de prévalence ou encore biais d'annotation.

Ces biais sont d'autant plus problématiques qu'ils sont difficiles à anticiper lors de la constitution des jeux de données et souvent complexes, voire impossibles, à corriger une fois le modèle entraîné.

Leur impact est similaire à celui du manque de généralisation : ils induisent un apprentissage partiel et biaisé de la réalité clinique, conduisant à des erreurs systématiques. Cela se traduit par une diminution des performances globales du modèle, notamment par une augmentation des faux positifs et des faux négatifs, ainsi qu'une baisse de l'aire sous la courbe ROC (AUC).

Au-delà de leur impact sur les performances, les biais soulèvent également des enjeux éthiques importants, en raison de l'hétérogénéité des résultats selon les populations. En effet, les biais n'affectent pas tous les patients de manière uniforme, et leurs effets varient selon les caractéristiques des individus analysés.

Cette variabilité peut conduire à des inégalités de prise en charge et à une forme de stigmatisation des populations sous-représentées dans les jeux de données. Par exemple, pour une pathologie identique, les enfants, les personnes âgées ou certaines minorités peuvent bénéficier de diagnostics moins précis et plus sujets à erreur que les populations majoritaires.

La maîtrise des biais constitue donc un enjeu majeur pour le développement d'une IA à la fois fiable et éthique. Les travaux actuels de recherche s'orientent notamment vers la constitution de jeux de données plus représentatifs des populations et des conditions cliniques réelles.

2.3. Les conséquences cliniques

Jusqu'à présent, les limites des systèmes d'IA appliqués à la radiologie ont été abordées principalement sous un angle technique. Toutefois, il est essentiel d'en comprendre les implications concrètes en pratique clinique, car toute limitation technique peut entraîner des conséquences organisationnelles, cliniques et éthiques significatives.

Les erreurs diagnostiques liées à une baisse de performance se traduisent notamment par une augmentation des faux positifs, c'est-à-dire des patients sans pathologie identifiés à tort comme malades. Cette situation est problématique, car elle conduit à la réalisation d'examens complémentaires parfois invasifs, contraignants ou inutiles (irradiation, hospitalisation, suivi infirmier rapproché).

Ces examens mobilisent également des ressources médicales et paramédicales, réduisant leur disponibilité pour d'autres patients. Par ailleurs, ils engendrent des coûts supplémentaires pour le système de santé (examens, produits de contraste, logistique) ainsi que pour les patients eux-mêmes (déplacements, reste à charge), dans un contexte de tension croissante sur les ressources médicales.

Il ne faut pas négliger non plus l'impact psychologique de ces parcours de soins, qui peuvent générer une anxiété importante et contribuer, dans certains cas, à une détérioration de l'état de santé global du patient.

À l'inverse, les faux négatifs, où une pathologie réelle n'est pas détectée, peuvent avoir des conséquences encore plus graves. Le diagnostic étant retardé, la prise en charge l'est

également, ce qui peut entraîner une aggravation de la maladie. Dans le cas de pathologies évolutives comme les cancers, la précocité du diagnostic est un facteur déterminant du pronostic. Un retard peut ainsi conduire à une perte de chance thérapeutique, voire à des issues fatales évitables.

Une autre limite importante concerne la confiance des professionnels de santé envers ces systèmes. Ce manque de confiance ne repose pas sur un rejet de principe, mais découle directement des limites évoquées précédemment. En effet, des performances élevées en phase d'entraînement ne garantissent pas une efficacité clinique réelle.

Par ailleurs, la variabilité des performances selon les jeux de données et la difficulté à garantir une stabilité des résultats rendent légitime une certaine prudence quant à leur utilisation. À cela s'ajoute le problème de l'opacité des modèles de deep learning, souvent qualifiés de modèles « boîte noire ». Cette absence d'explicabilité limite la compréhension des décisions prises par l'algorithme et complique la validation clinique des résultats.

Ainsi, l'impossibilité d'expliquer le raisonnement d'un modèle constitue un frein majeur à son acceptation, car elle empêche toute vérification des fondements d'un diagnostic.

En conclusion, les systèmes d'IA appliqués à la radiologie présentent encore des limites importantes, qu'elles soient techniques, éthiques, cliniques ou liées à la confiance des professionnels de santé. Toutefois, ces limites ne remettent pas en cause leur potentiel, mais soulignent la nécessité de poursuivre les efforts de recherche afin de surmonter ces obstacles et de permettre une intégration sécurisée et efficace de ces outils en pratique clinique.

III. Les perspectives d'amélioration pour développer des systèmes plus robustes et généralisables

Après avoir défini les caractéristiques que devrait présenter un modèle d'IA idéal pour l'analyse radiographique, ainsi que les principaux obstacles à sa conception, il convient d'examiner les solutions actuellement proposées par la littérature scientifique pour s'en rapprocher. En effet, toutes les limites identifiées ne disposent pas nécessairement de solutions immédiatement applicables en pratique.

3.1. Améliorer les jeux de données

Depuis le changement de paradigme plaçant la qualité des données au centre des performances des algorithmes, les jeux de données sont désormais considérés comme un facteur déterminant dans la conception d'une IA efficace en radiologie clinique. Il a également été montré qu'ils doivent être représentatifs afin de garantir des résultats fiables

et non biaisés. Toutefois, cela ne précise pas les leviers concrets permettant d'améliorer cette représentativité.

Kelly et Kim, dans leur revue systématique, identifient plusieurs axes d'amélioration. Le premier consiste à augmenter la taille des jeux de données en incluant davantage de patients. Cela permet d'intégrer des cas rares et d'améliorer la capacité du modèle à gérer des situations atypiques.

Un autre levier majeur repose sur la diversification des sources de données. Il s'agit notamment de s'affranchir des contraintes liées aux pratiques locales d'acquisition et de traitement des images, en multipliant les centres hospitaliers participants et en utilisant différents équipements d'imagerie.

Enfin, un grand volume de données ne garantit pas à lui seul la diversité. Il est donc essentiel d'assurer une représentation équilibrée des populations sur les plans générationnel, ethnique et géographique. Cette diversité peut être renforcée par des collaborations internationales favorisant le partage de données.

L'ensemble de ces mesures permet d'obtenir des jeux de données plus représentatifs de la réalité clinique et améliore ainsi la capacité de généralisation des modèles.

Au-delà de la taille et de la diversité, la qualité des annotations constitue également un élément fondamental. Les annotations représentent la référence permettant d'associer une image à un diagnostic. Plus elles sont précises et fiables, plus l'apprentissage du modèle est pertinent. À l'inverse, des annotations incomplètes ou erronées entraînent des apprentissages biaisés.

Pour être optimales, les annotations devraient être validées par plusieurs radiologues afin d'aboutir à un consensus et de limiter les erreurs. L'intégration d'une notion d'incertitude dans les annotations peut également constituer une information utile pour le modèle, lui permettant de hiérarchiser les données d'apprentissage.

La question des biais reste également centrale. Comme évoqué précédemment, leur particularité réside dans leur difficulté d'identification et de correction une fois le modèle entraîné. Il est donc crucial d'agir en amont, lors de la constitution des jeux de données, en veillant à l'équilibre des classes et à la représentativité des populations.

La documentation des jeux de données constitue également un levier important, car elle permet de mieux comprendre leur composition et de limiter certains biais évitables. Enfin, des audits réguliers sont nécessaires afin de s'assurer que les jeux de données restent adaptés à la population cible, celle-ci évoluant dans le temps.

Ainsi, l'amélioration des jeux de données ne repose pas uniquement sur leur taille, mais sur une approche globale intégrant représentativité, qualité des annotations, maîtrise des biais et actualisation continue.

3.2. Améliorer les méthodes de développement et d'évaluation des modèles

Malgré l'évolution des pratiques, la littérature scientifique continue d'accorder une importance majeure à l'amélioration des méthodes de développement et d'évaluation des modèles.

L'une des recommandations principales est la généralisation de la validation externe pour les modèles de deep learning en radiologie. Kim et Kelly soulignent que cette étape permet d'évaluer les performances d'un modèle sur des données indépendantes, reproduisant ainsi les conditions proches de la pratique clinique.

Ils identifient quatre critères essentiels pour une évaluation robuste : tester le modèle dans plusieurs pays, sur différentes populations, dans plusieurs établissements et sur différents équipements d'imagerie. Cette approche réduit le risque de valider des modèles incapables de généraliser.

La standardisation des études constitue un autre enjeu majeur. Afin de comparer objectivement les performances des algorithmes, il est nécessaire de disposer de protocoles méthodologiques rigoureux et homogènes. Dans ce cadre, deux cadres de référence sont particulièrement utilisés : SPIRIT-AI, qui encadre la conception des essais cliniques en intelligence artificielle, et CONSORT-AI, qui vise à standardiser la présentation des résultats afin de garantir leur transparence, leur reproductibilité et leur comparabilité.

Enfin, l'évaluation des modèles ne doit pas se limiter à la phase de conception. Les données cliniques, les pratiques médicales et les populations évoluent en permanence. Les modèles doivent donc être régulièrement réentraînés, recalibrés et mis à jour afin de rester pertinents dans le temps.

Une surveillance continue des performances est également indispensable afin de détecter toute dégradation potentielle liée à des changements de contexte.

En définitive, la qualité d'une IA ne dépend pas uniquement de ses performances initiales, mais également de la rigueur de son évaluation et de sa capacité à être suivie dans le temps. Les cadres SPIRIT-AI et CONSORT-AI illustrent cette exigence de standardisation et de transparence. Toutefois, ces principes doivent être maintenus après le déploiement clinique, car de bonnes performances initiales ne garantissent pas la stabilité du modèle dans des conditions évolutives.

3.3. Favoriser l'intégration clinique des systèmes d'intelligence artificielle

Les améliorations présentées jusqu'ici visent à développer des systèmes d'intelligence artificielle plus robustes, fiables et généralisables. Toutefois, ces avancées techniques ne garantissent pas à elles seules leur intégration en pratique clinique.

Un modèle d'IA doit avant tout démontrer un bénéfice concret dans la prise en charge des patients. Les performances statistiques doivent ainsi se traduire par une amélioration du diagnostic, une prise en charge plus rapide, une réduction de la charge de travail des radiologues et, plus globalement, une amélioration du parcours de soins.

Une autre limite importante concerne l'explicabilité des modèles. Les systèmes de deep learning sont encore largement considérés comme des « boîtes noires », dont le fonctionnement interne reste difficile à interpréter. Cette opacité constitue un frein majeur à leur adoption clinique, car elle limite la compréhension et la validation des décisions produites.

Dans ce contexte, le domaine de l'*Explainable AI* (XAI) prend une importance croissante. Certaines méthodes permettent aujourd'hui de mieux interpréter les décisions des modèles. En radiologie, des approches comme Grad-CAM permettent de générer des cartes de chaleur mettant en évidence les zones de l'image ayant le plus influencé la prédiction.

Ces visualisations peuvent être superposées aux images radiographiques, facilitant ainsi l'analyse du raisonnement de l'algorithme par le radiologue et permettant de détecter d'éventuelles erreurs ou biais.

Enfin, il est essentiel de replacer l'IA dans son rôle réel. Contrairement à certaines perceptions, l'IA ne vise pas à remplacer les radiologues, mais à les assister. Elle doit être envisagée comme un outil d'aide au diagnostic ou un second lecteur.

Les approches récentes convergent vers un modèle de collaboration homme-machine, combinant les forces complémentaires des deux acteurs : l'expertise clinique, le raisonnement et la prise en compte du contexte pour l'humain, et la rapidité d'analyse ainsi que la capacité de traitement de grands volumes de données pour l'IA.

Cette complémentarité apparaît aujourd'hui comme la stratégie la plus prometteuse pour une intégration durable et efficace de l'intelligence artificielle en radiologie.

Conclusion

Cette revue de littérature met en évidence que les caractéristiques des jeux de données constituent aujourd'hui l'un des principaux déterminants des performances des systèmes d'intelligence artificielle appliqués à la radiographie. La qualité des annotations, la diversité des populations représentées, le caractère multicentrique des données, ainsi que la prise en compte de l'incertitude diagnostique influencent directement la capacité des modèles à apprendre des informations pertinentes, à limiter les biais et à généraliser leurs performances en pratique clinique. Ainsi, les jeux de données ne représentent plus un simple support d'entraînement, mais un véritable levier de développement des systèmes d'intelligence artificielle.

Toutefois, cette revue montre également que la qualité des jeux de données ne constitue pas, à elle seule, une garantie de performance clinique. Malgré les progrès considérables réalisés ces dernières années, les modèles de deep learning restent confrontés à plusieurs limites majeures, parmi lesquelles le manque de généralisation, les biais liés aux données et les difficultés d'intégration dans des environnements cliniques variés. Ces limites peuvent entraîner une diminution des performances diagnostiques, mais également soulever des enjeux éthiques en raison d'inégalités potentielles entre les populations étudiées.

Face à ces constats, les travaux de recherche s'orientent désormais vers une approche davantage centrée sur les données (*data-centric AI*). L'amélioration de la représentativité des jeux de données, le développement de procédures de validation externe multicentriques, la standardisation des méthodes d'évaluation, ainsi que le développement d'une intelligence artificielle plus explicable apparaissent aujourd'hui comme les principaux leviers permettant de concevoir des systèmes plus robustes et plus fiables. Ces évolutions visent non seulement à améliorer les performances des modèles, mais également à renforcer la confiance des professionnels de santé et à favoriser leur intégration dans la pratique clinique.

Enfin, l'ensemble des travaux analysés converge vers une même conclusion : l'intelligence artificielle n'a pas vocation à remplacer le radiologue, mais à l'accompagner dans sa pratique quotidienne. Son intérêt réside dans sa capacité à assister le professionnel de santé en améliorant la détection de certaines anomalies, en facilitant la priorisation des examens ou en apportant une aide à la décision. L'avenir de l'intelligence artificielle en radiologie dépendra donc moins des seules avancées algorithmiques que de notre capacité à concevoir des jeux de données de haute qualité, représentatifs de la diversité des patients et des pratiques cliniques, afin de garantir des systèmes performants, généralisables et véritablement bénéfiques pour les patients.

Références Bibliographiques

I. Caractéristiques des jeux de données et IA en imagerie médicale

1. Arvaniti E, Claassen M. Dataset quality in medical imaging artificial intelligence. PMC. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6268174/>
2. Seyyed-Kalantari L, et al. Bias and generalization in medical imaging datasets for deep learning. PMC. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8017414/>
3. Domingos P. The role of data in artificial intelligence systems. AAAI Conference Proceedings. Available from: <https://ojs.aaai.org/index.php/AAAI/article/view/3834>
4. Zhang Y, et al. Generalization in deep learning: theoretical insights and practical implications. arXiv. 2021. Available from: <https://arxiv.org/abs/2105.03020>

II. Limites actuelles des jeux de données en radiologie

5. van der Putten L, et al. Domain shift in medical imaging artificial intelligence: a systematic review. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9152694/>
6. Draelos RL, et al. Dataset bias in clinical machine learning applications. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12689012/>
7. Rajpurkar P, et al. Challenges in deploying artificial intelligence in medical imaging. *ScienceDirect*. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S1546144023005070>
8. Oakden-Rayner L, et al. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn*. 2020;151-159. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8674135/>
9. Chen IY, et al. Bias and fairness in healthcare artificial intelligence systems. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9650160/>
10. Obermeyer Z, et al. Clinical consequences of algorithmic bias in healthcare. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9931338/>
11. Holzinger A, et al. Explainability and trust in medical artificial intelligence. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9668941/>

III. Perspectives d'amélioration des systèmes d'IA

12. Chen RJ, et al. Improving dataset quality and diversity in medical imaging AI. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10698597/>
13. Draisbach C, et al. External validation of AI models in medical imaging. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9533091/>
14. Liu X, et al. SPIRIT-AI and CONSORT-AI guidelines for clinical trials in artificial intelligence. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9530765/>
15. van der Velden BHM, et al. Explainable artificial intelligence in medical imaging: a review. *PMC*. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11491635/>