UNIVERSITÉ DE LILLE

FACULTE DE MÉDECINE HENRI WAREMBOURG

Année : 2020

THÈSE POUR LE DIPLOME D'ÉTAT

DE DOCTEUR EN MÉDECINE

Intelligence artificielle et dépistage du cancer du sein : Étude de performance de deux logiciels

Présentée et soutenue publiquement le 9 avril à 18h
au Pôle Formation

par Adrien LE VOURCH

_____

JURY

Président :

Monsieur le Professeur Philippe Puech

Assesseurs :

Monsieur le Professeur Olivier Ernst

Monsieur le Professeur Emmanuel Chazard

Directeur de thèse :

Monsieur le Docteur Nicolas Laurent

_____

ABREVIATIONS

BI-RADS	Breast Imaging – Reporting and Data System
IA	Intelligence artificielle
TVP	Taux de Vrai Positif
TFP	Taux de Faux Positifs
ROC	Receiving Operater Chart
MIT	Massachussets Institute of Technology
MS	Mammographie de synthèse
MD	Mammographie digitale
ROC	Receiving Operator Characteristic
AUC	Area Under the Curve

Sommaire

INTRODUCTION

1 INTELLIGENCE ARTIFICIELLE, DATA ET ALGORITHMES

Si l'intelligence artificielle (IA) est un concept connu depuis des années[1], l'augmentation récente des capacités de calcul algorithmique des ordinateurs a rendu possible leur développement et leur utilisation dans des domaines qui touchent déjà notre quotidien[2].

Dans le domaine médical, de nombreuses applications se développent dans l'ensemble des spécialités, que ce soit en dermatologie, en cancérologie ou en opthalmologie[3]. Mais la spécialité dont les habitudes risquent d'être le plus bouleversées est probablement l'imagerie médicale, en raison de l'informatisation et de la facilité de collecte des données ou data nécessaire à son fonctionnement.

Si les radiologues sont conscients des avancées technologiques récentes en la matière, de nombreux praticiens restent circonspects devant les futures applications et la place des outils dans leur pratique quotidienne, avec une part de méfiance envers une technologie promise à les aider dans leur travail où même à les remplacer[4].

Une étude récente portant sur un questionnaire soumis à un panel de 270 radiologues de tout horizon, internes, praticiens hospitaliers et privés, a montré que ces derniers disposaient de peu d'informations et de connaissances sur le sujet mais étaient en grande majorité volontaires pour se former[5].

1.1 Sémantique: IA, Machine Learning, Deep learning?

Il apparaît désormais évident que la formation du radiologue devra passer par la compréhension du langage et du fonctionnement des algorithmes, au même titre que l'apprentissage des techniques scanner ou IRM.

L'IA est une branche des Sciences de l'Ordinateur ou Computer Science, visant à imiter le raisonnement humain, dans des domaines vastes comme la robotique, la recherche, le raisonnement ou le langage.

Le Machine Learning (ML) est une des composantes de l'IA, consistant à entraîner un algorithme à la réalisation d'une tâche grâce à un jeu de données existant. Il existe de

nombreuses techniques de ML en fonction de l'analyse que l'on désire effectuer : Classification, Régression, Segmentation…[6,7]

Les réseaux de Neurone Artificiels ou Artificial Neural Networks (ANN), sont une des techniques de ML particulièrement utilisée en radiologie, dont le fonctionnement peut être rapporté de façon lointaine à celui des neurones biologiques. L'information transitant à travers le réseau affecte la structure de l'ANN en fonction de l'entrée et de la sortie des données, permettant une adaptation ou un apprentissage en fonction du jeu de données.

Une autre façon de voir la chose serait d'imaginer, dans une constellation d'image, les algorithmes capables de tracer une frontière géométrique entre les objets selon leurs caractéristiques[8].

Le terme de Deep Learning (DL) est utilisé lorsque ces réseaux neuronaux sont organisés en de multiples couches (layer), plus ou moins profondes interagissant entre elles. Les différentes couches correspondent à des unités de traitement non linéaires pour l'extraction et le traitement des données sur plusieurs niveaux[7].

Si la performance de ces logiciels est en constante amélioration, les paramètres utilisés pour arriver aux résultats demeurent dans la plupart des cas inconnus. Le terme de "boites noires" ou "black box" est couramment utilisé pour désigner le manque de connaissances sur les moyens mis en œuvre par l'algorithme pour arriver au résultat.

1.2 Data et Algorithmes

La première étape de la construction d'un logiciel est le choix de l'algorithme. La majorité des développeurs ont recours à un catalogue d'algorithme « open source » (gratuits) comme base. Différents types de structures algorithmiques sont à disposition selon la fonction souhaitée de l'algorithme.

Pour travailler l'algorithme, différents outils, appelés « Framework », existent et permettront son modelage de via différents langages mathématiques. Python est le plus utilisé dans ce domaine.

Les Framework les plus utilisés sont distribués par Google, Facebook ou encore Amazon.

Les données labelisées correspondent aux données bénéficiant de l'ajout d'annotations qui permettront à l'algorithme de bénéficier d'un entraînement adéquat. Il s'agit par exemple d'ajouter le nom de certains objets ou de cibler ces derniers sur des photos

Pour être performant, les algorithmes de ML ont besoin d'un entraînement. Cet entraînement peut être « supervisé » par des données labellisées, ou « non supervisé ». La collecte et l'annotation des données labellisées constituent une part majeure du développement.

Les algorithmes « non supervisés » vont trouver d'eux même les similitudes entre les images non labellisées désignant explicitement la lésion. Ce type d'algorithme est particulièrement adapté par exemple à la Radiomic, étudiant les textures et les nuances d'une image invisible à l'œil nu.

La plupart des algorithmes utilisés en imagerie sont « supervisés », chaque image est associée au critère Pathologique ou Normal. Chaque image est également labellisée avec par exemple, en sénologie, un contourage de la lésion et un label indiquant le classement de la lésion selon le BI-RADS et donc sa probabilité de malignité.

Enfin il existe un entraînement hybride dit semi-supervisé ou les images sont labellisées par exemple en fonction de leur caractère pathologique ou non mais sans que la lésion soit explicitement contourée[7].

La plupart des logiciels aujourd'hui utilisent différents types d'algorithmes superposés au sein de leurs structures afin de mener à bien leurs tâches, ou même en amont de leur exécution, afin de trier ou récupérer les données.

1.2.1 Archivage des données

Les données de santé stockées chaque année sont colossales, et leur volume a été multiplié par 9 entre 2016 et 2018, une hospitalisation classique générant environs 150 000 pièces de données[9].

Les data en radiologie ont l'avantage d'être archivées depuis la généralisation des PACS (Picture Archiving and Communication System) en France il y a une dizaine d'année[10].

Les données archivées au format Dicom comprennent en plus des images toute une série d'informations allant des machines utilisées, des paramètres techniques, de la dose reçue par le patient, jusqu'au compte rendu, voir à des informations cliniques. La

standardisation du format de stockage a été un moteur pour l'utilisation de ces données.
L'évolution des algorithmes amènera à les entraîner avec de plus en plus d'informations, notamment des informations cliniques utiles pour une prise en charge globale. Si le stockage de ces informations au sein des fichiers Dicom peut s'avérer sensible en raison des informations contenues, l'amélioration de la sécurisation des données et une vigilance accrue permettra l'amélioration de ces bases de données.
Si l'ensemble des centres radiologiques disposent d'une banque de données informatisées notamment sur le système PACS, ces données dans nos applications ne peuvent être utilisées à l'état brut pour entraîner les algorithmes « supervisés ». Une grande partie du travail consiste donc à labelliser les images.

Pour trier les images, si les comptes rendus ont été informatisés, d'autres logiciels de ML peuvent venir en aide. Ces algorithmes, appelés Natural Processing Language, vont être capables de trier les études en fonction de mots clés contenus dans les comptes rendus.
La deuxième étape de l'annotation va consister à poser des labels sur ces images afin que le logiciel puisse s'entraîner efficacement : chaque lésion sera contourée manuellement ou semi-manuellement par un opérateur expert. Cette étape est capitale car les performances de l'algorithme vont dépendre directement de la qualité de labellisation des images.
Les radiologues qui réalisent les annotations des data le font dans le cadre de partenariats ou sont rémunérés à l'acte en fonction des accords.
Les images utilisées par les algorithmes doivent être transformées pour l'analyse et la résolution abaissée afin de diminuer le temps et les ressources nécessaires aux logiciels pour fonctionner. Les images sont principalement travaillées sur un format .jpeg et non .raw pour les mêmes raisons. L'impact de la baisse de résolution n'a à ma connaissance jamais été évaluée pour ces logiciels.

1.2.2 Valorisation des données

Si tous les centres disposent d'un certain nombre de données sur leurs PACS, la valorisation de ces data n'est pas quelque chose d'aisée et les radiologues exerçants

dans les structures publiques comme privées essaient de s'organiser pour essayer de se servir au mieux de ces données.

A la pratique du partenariat d'un centre local avec un des acteurs du marché s'oppose une vision plus globale avec une mise en commun de ces bases de données. Mais la France n'est pas en avance, et les grands projets ont mis du temps à être opérationnels, notamment en raison des formalités administratives lourdes et des systèmes informatiques indépendants, propre à chaque centre hospitalier.

A l'heure actuelle plusieurs projets de mutualisation sont plus ou moins avancés dans leur développement.

Concernant les données de santé générales, le Data Health Hub se voit comme « un catalogue de bases de données les plus prometteuses » mis à disposition des chercheurs, des associations de patients ou encore des start-ups.

Le projet DRIM IA sur le point de voir le jour se définit comme un outil pour les radiologues, au service du patient, sous la forme d'une plateforme numérique connectant aussi bien les cabinets libéraux que les structures hospitalières et les CHU[11]. Il fonctionnera comme un outil de standardisation et d'indexation des données, afin de servir la recherche et le développement d'outil par les industriels.

Les données seront envoyées par les structures à cette plateforme numérique, qui s'occupera de leur anonymisation et de leur sécurisation, avant la mise en relation avec des industriels qui pourront utiliser les data en échange de rémunération ou de services.

L'Assistance Publique des Hôpitaux de Paris a présenté également son "Entrepôt de donnée" ou Data Warehouse, avec son propre système de stockage et d'archivage des données en vue d'une utilisation en recherche clinique ou par l'industrie, avec laquelle le groupement hospitalier a déjà noué des partenariats. Un projet de recherche spécifique à la mammographie a également vu le jour au sein de la structure en partenariat avec les équipes de recherche de l'Institut des Sciences du Calcul de Données (ISCD), nommé EZ Mammo, et a pour but la construction d'une base de données afin de tester et de valider de façon externe et indépendante les algorithmes de détection et d'évaluation de la densité mammaire.

Par ailleurs de multiples structures ont noué des partenariats à l'instar de Valenciennes avec des industriels, avec d'un côté le partage des données avec le développeur de logiciel pour l'entraînement des logiciels et de l'autre la possibilité de co-développer et utiliser des algorithmes.

Aux États-Unis, plusieurs projets tâchant de mettre en lien directement les patients avec les industriels afin de monétiser leurs data, ont vu le jour. Ils fonctionnent et sont sécurisés grâce à la technologie de la Blockchain, utilisée notamment dans les cryptomonnaies. Le site du chercheur Dexter Hadley https://www.breastwecan.org/, qui proposait aux femmes de mettre directement leurs examens sur sa base de données est aujourd'hui inactif.

1.2.3 Cadre éthique et juridique

Une réglementation et un cadre permettant la recherche et l'innovation apparaissent essentiels, pour éviter un risque de blocage du développement de nouveaux logiciels, et d'importer des solutions développées à l'étranger dans un cadre éthique non contrôlé.

Le premier cadre éthique relève de la déontologie médicale et des règles juridiques qui encadrent la profession[12].

Les progrès de l'IA et de ses algorithmes sont allés plus vite que la réflexion sur la partie éthique et réglementaire l'entourant.

La balance bénéfice / risque est difficilement établie avec l'IA, en raison d'une part inconnue du fonctionnement des algorithmes (black box), et des risques liés au traitement de données personnelles et confidentielles non quantifiables[13].

Un exemple démonstratif concerne les logiciels des voitures à conduite autonome : récemment une personne est décédée aux Etats-Unis lors d'un accident impliquant un de ces véhicules. Si pour l'instant les voitures autonomes ont pu démontrer leur fiabilité avec une baisse drastique du nombre d'accidents à kilométrage égal par rapport à un opérateur humain, il s'agit toujours de définir une responsabilité en cas d'accidents. Doit-on sanctionner le fabricant du logiciel ? La personne dans la voiture ? Est-ce une faute intrinsèque du logiciel ?

De plus, lorsqu'il s'agit de faire des choix cruciaux, l'absence de connaissances de l'ensemble des facteurs régissant l'algorithme peut soulever des questions éthiques : comment être sûr que l'algorithme ne se base pas sur des considérations ethniques ou sociales ?

L'université du Massachussetts Institute of Technology (MIT) a développé une étude se basant sur un site internet en libre accès proposant aux internautes du monde entier de choisir entre des vies humaines à la place d'un logiciel de voiture autonome: rouler sur des piétons, ou accidenter les passagers du véhicule dans diverses situations, en fonction du sexe, du morphotype, de l'âge et de la classe socio-économique des victimes. Si l'ensemble des participants ont choisi d'épargner le plus de vies humaines sur chacune des situations, il est apparu de nombreuses disparités entre les pays et les régions du globe, chaque culture détenant sa propre vision de l'éthique.

Il apparaît clairement que les législations devront se faire non pas à l'échelle des logiciels mais à l'échelle des pays[14].

En Europe et notamment en France, le RGPD (Règlement Général sur la Protection des Données) pose les bases de la législation concernant le respect de la vie privée, la propriété des données.

Car si jusqu'à présent les données étaient utilisées dans le cadre de la recherche, dans l'intérêt du patient, le développement des solutions d'IA comme des produits commerciaux amènent à donner aux examens et bases de données une valeur mercantile nouvelle.

Le RGPD donne également des éléments de réponse concernant la propriété des données : le patient est propriétaire de ses données et peut choisir à tout moment de retirer son consentement. Mais les données se conçoivent avec d'une part les images qui sont la propriété du patient et d'autre part le compte rendu et les labels, pour lesquels il peut se concevoir que ceux-ci appartiennent au radiologue qui a fourni l'acte.

De ce fait si des données doivent être utilisées pour la recherche ou vendues à des entreprises, le patient doit être informé de l'usage qui va en être fait et pourrait donc en théorie donner son accord pour une utilisation universitaire mais pas mercantile[15].

En Europe, pour effectuer des recherches sur des données anonymisées et rétrospectives recueillies de façon habituelles sans recueil « exprès » du consentement, une homologation de la méthodologie de référence MR003 ou 004 auprès du CNIL est nécessaire en fonction de l'implication ou non de la personne humaine. Si la personne humaine est impliquée, un avis favorable du CPP sera nécessaire avant le début de la recherche. Les patients doivent néanmoins être informés de façon générale avec par exemple des affiches dans le service concernant les travaux de recherche, ainsi qu'une information individuelle sur les projets en cours.

Les médecins doivent informer les patients sur les modalités du droit d'accès, de rectification et d'opposition via des documents remis au patient ou la signature d'une convention signée par les professionnels de santé[16]. Les données peuvent ensuite être utilisées et conservées jusqu'à la mise sur le marché du produit concerné. Seul le professionnel de santé dirigeant la recherche est responsable de l'accès aux données anonymisées et il est également le seul à disposer de la possibilité de croisement entre les données d'anonymisation et le nom et prénom du patient[17].

1.2.4 Intégration des logiciels à l'environnement de travail

L'intégration des logiciels d'IA au workflow et aux consoles de lecture est primordiale pour une utilisation fluide et un gain de temps et d'efficacité dans la pratique quotidienne.

Les premiers logiciels disponibles maintenant et dans un futur proche contiennent toute une palette d'outils susceptibles de nous faire gagner du temps, allant de la segmentation automatique pour une volumétrie à la rédaction de pré comptes-rendus, les radiologues vont devoir se poser la question de leur place dans leur activité.

L'utilisation de l'IA peut se faire "à la demande", sur une image ou une série d'image en fonction de la modalité utilisée d'imagerie, pose problème.

L'intérêt de cette méthode est de diminuer les faux positifs car seules les images suspectes visibles par le radiologue vont être soumises à l'interrogation de l'algorithme. L'inconvénient réside dans de potentiels faux-négatifs, car seules les images analysées comme suspectes par le radiologue seront soumises au logiciel.

Une autre solution réside dans l'envoi automatique des données pour un pré-traitement des données par l'algorithme qui pourrait ensuite prioriser les études à valider par le radiologue via un système de flagging. Ce système semble peu adapté en France car les examens d'imagerie sont interprétés rapidement après leur acquisition.

L'envoi automatique des données sur le RIS après le post traitement est la forme la plus autonomisée du fonctionnement des logiciels d'IA, et nécessite une gestion des données où le radiologue pourrait être en contradiction avec l'algorithme. Ce système permettrait notamment dans les services d'urgence d'avoir des résultats de façon extrêmement rapide mais nécessiterait d'avoir un algorithme extrêmement fiable d'une part, et d'autre part un système de gestion des données validées par le radiologue

pour la gestion des faux positifs et des faux négatifs. Il est peu probable de voir se développer rapidement ce genre de solution, intégrant la valeur ajoutée du radiologue, dans un pays comme la France ou l'accès à l'imagerie notamment en urgence reste satisfaisante[18].

1.2.5 Temps de lecture

Un des avantages des logiciels avancés dans les différentes études est le gain de temps procuré par les logiciels sur la lecture des mammographies. Ce gain de temps doit être pondéré par la prise en main du logiciel et la courbe d'apprentissage nécessaire à sa bonne utilisation.

La plupart des logiciels, s'ils sont reliés au workflow, sont accessibles sur une console séparée de la console d'interprétation, le temps d'interprétation étant peu conséquent pour la lecture d'une mammographie, il est peu probable de voir une amélioration significative de cette composante grâce à l'IA.

Le temps gagné pourrait concerner les composantes techniques ou aider à la pré-rédaction d'un compte rendu dans le domaine de la sénologie ou de la radiographie standard.

Mais le gain de temps n'est pas le but ultime des algorithmes, et si l'amélioration visible du taux de faux positifs et faux négatifs doit passer par des algorithmes qui prennent légèrement plus de temps que lors d'un examen de routine, l'objectif d'amélioration du service rendu aux patientes sera atteint.

1.2.6 IA, expertise et formation

L'apport de l'IA en imagerie médicale à pour objectif de simplifier l'interprétation des examens d'imagerie, de fournir un soutien au radiologue notamment dans la réalisation de mesures, dans la pré-rédaction de comptes rendus, dans le ciblage de lésions suspectes.

Il est par ailleurs primordial que le radiologue garde son expertise et son analyse critique.

Comment former les futurs internes à l'analyse et au traitement des données de façon experte si les logiciels de demain explorent, segmentent et analysent les données de façon semi-autonome ?

La formation médicale va devoir s'adapter à l'évolution des méthodes de travail.

Par ailleurs la création pour les logiciels de gigantesques bases de données standardisées et annotées pourrait également servir à la formation et l'entraînement des internes et des radiologues pour peu que des solutions didactiques se développent.
Une des pistes avancées par une faculté australienne serait l'entraînement des radiologues grâce à des images crées par des algorithme d'IA dits "réseaux de neurones adverses génératifs"(Générative Adversarial Network - GAN), permettant de créer des images aléatoires à partir d'un set d'images prédéfini[19].
Ces GAN pourraient également participer à l'enrichissement artificiel de sets de données sur des pathologies rares (en taille et en hétérogénéité), ainsi qu'à l'entraînement des algorithmes de reconnaissance d'images.
Les principaux obstacles aujourd'hui sont la création d'images haute résolution et l'absence de concordance des images à des marqueurs biocliniques[20].

2 Le dépistage

2.1.1 Système BI-RADS et examens mammographiques

L'examen mammographique dans le cadre du dépistage du cancer du sein est composé de 4 incidences selon les recommandations officielles, R CC, R MLO, L CC, L MLO, correspondant aux vue droites et gauche du sein de face et en oblique médio-latéral. Sont ajoutées dans plus en plus de centres des vues en tomosynthèse de face bilatérale.
Le compte-rendu radiologique est basé sur le système de classification de référence BI-RADS, apparu vers la fin des années 1990 aux États-Unis pour standardiser l'interprétation des radiologues. Le système a subi plusieurs modifications depuis son introduction. La version la plus récente est :

- BI-RADS 1 : Examen négatif : mammographie normale.
- BI-RADS 2 : Constatations bénignes.
- BI-RADS 3 : Anomalie probablement bénigne (>98%). Surveillance à court terme proposée.

- BI-RADS 4 : Anomalie suspecte, une biopsie doit être envisagée.
- BI-RADS 5 : Haute probabilité de malignité, une action appropriée doit être entreprise.

2.1.2 Le dépistage organisé du cancer du sein en France

Depuis 2004, le dépistage du cancer du sein chez la femme est généralisé à tout le territoire et concerne toutes les femmes de 50 à 74 ans, avec la réalisation de mammographies deux incidences, face et oblique externe, bilatérales, tous les deux ans. Sont exclues de ce dépistage de façon temporaire toutes les femmes pour lesquelles une anomalie bénigne a été détectée et pour laquelle une surveillance de cette lésion est établie. Sont exclues de façon définitive, toutes les femmes ayant eu ou étant en cours de traitement pour un cancer du sein, les femmes ayant des facteurs de risque avérés de cancer du sein, avec des mutation génétiques particulières ou ayant déjà bénéficié d'une chirurgie pour une lésion dont le type histologique est à risque[21].

Une particularité du dépistage en France est la réalisation d'un examen clinique systématique et d'un bilan de diagnostic immédiat suite à la mammographie en cas d'anomalie clinique ou mammographique.

Pour toutes les études de dépistage Bi-RADS 1 et 2, les clichés sont envoyés par la poste en seconde lecture avec les clichés antérieurs. En cas de discordance entre le premier et le deuxième lecteur, l'information est transmise par le centre de dépistage qui contactera la patiente pour un examen auprès du radiologue qui avait réalisé la mammographie pour compléter les explorations.

Les études Bi-RADS 3, 4, 5 mammographiques ne sont pas envoyées pour une deuxième lecture car elles nécessitent une surveillance rapprochée ou un examen complémentaire, type échographie, IRM ou biopsie mammaire.

Pour que les radiologues participent au dépistage en tant que premiers lecteurs, ils ont une obligation de lecture d'au moins 500 mammographies par an, portée à 1500 pour les seconds lecteurs.

La double lecture est instaurée en France depuis 2012. Il a en effet été montré que dans le dépistage organisé du cancer du sein réalisé par l'HAS, 9% des cancers du

sein avaient été dépistés en seconde lecture lors une campagne dont la prévalence était de 7 à 8 pour 1000 [22]. L'efficacité de la double lecture augmente quand la détection est difficile, comme lors d'une première participation à un dépistage en l'absence d'examen antérieur, chez une patiente présentant des seins denses ou avec une lésion de petite taille.
Mais la double lecture augmente également le temps, les ressources et le coût du dépistage. Si plusieurs études ont prouvé l'efficacité de la double lecture en termes de coût-efficacité, ces études sont anciennes et ne concernaient pas les mammographies digitales et l'intégration de la tomosynthèse en dépistage.

2.1.3 Le dépistage décentralisé aux Etats-Unis

Il n'existe pas de dépistage organisé centralisé du cancer du sein aux États-Unis, les patientes doivent effectuer elles même la démarche auprès de leur médecin de proximité afin de réaliser un examen de dépistage. L'information pour encourager les patientes à se faire dépister est délivrée par des campagnes d'information financées par les autorités de santé publique, les organismes d'assurance divers, mais aussi par la publicité dispensée par les centres radiologiques eux-mêmes, pratique impensable et interdite dans un pays comme la France qui dispose d'une couverture de santé universelle et gratuite. Le prix des mammographies est bien plus important qu'en France, avec un coût moyen par état oscillant entre 200$ et 400$[23], contre 66,42€ en France. La prise en charge du coût des mammographies dépend de plusieurs entités publiques ou privées en fonction de l'âge de la patiente, avec une couverture dédiée après 65 ans, du statut de travailleur de la femme ou de son conjoint ainsi que de ses ressources économiques.

L'absence d'organisme de santé publique de référence dispensant les recommandations rend impossible l'harmonisation des pratiques. Si la United States Preventive Services Task Force recommande une mammographie de dépistage tous les deux ans entre 50 et 74 ans, d'autres sociétés savantes comme l'American Cancer Society font état d'une mammographie annuelle ou biannuelle à partir de 40 ans[24]. En 2013 plus de la moitié des femmes âgées de 40 à 50 ans aux États-Unis avaient bénéficié d'une mammographie annuelle[25]. En pratique les femmes réalisent leur examen dans un centre de radiologie, avec les 4 vues de base : CC et MLO pour

chaque côté. La lecture est dite différée, car les études sont lues à postériori par les radiologues américains lors de vacations dédiées. En cas de lésion suspecte ou de doute, les femmes sont rappelées dans un processus nommé « recall » pour la réalisation d'investigations complémentaires. De plus il n'existe pas de deuxième lecture en dehors d'initiatives locales.

Ce marché est donc économiquement capital pour les start-ups et leurs investisseurs, ces derniers recherchant avant tout un retour sur investissement rapide.

2.1.4 Stratification du risque de développer un cancer du sein

Concernant le dépistage individuel, il existe des recommandations pour les femmes à haut risque, mais aucune recommandation pour les femmes à risque intermédiaire.
Si le dépistage est aujourd'hui ancré dans les territoires et les mentalités, son organisation et son fonctionnement est susceptible d'être chamboulés dans les années à venir.
Les logiciels d'IA peuvent changer la notion de stratification du risque en fonction de la densité, des antécédents familiaux et génétiques, de l'âge et des autres facteurs de risque, afin de créer un score de risque personnalisé pour les patientes.
En fonction de ce score, l'approche et la durée entre deux mammographies pourront être allongées ou raccourcies.
Plusieurs études sont en cours dans le monde, dont une étude européenne, nommée MyPeBS, dont l'objectif est de comparer l'approche classique du dépistage du cancer du sein à une approche personnalisée. Cette dernière devrait permettre d'adapter la fréquence des mammographies en fonction du risque de développer un cancer calculé en fonction de multiples paramètres, dont des facteurs génétiques à l'aide d'un test salivaire.
Le centre Hospitalier de Valenciennes fait partie des centres d'inclusions en France.
Aux États-Unis, une étude a montré qu'un algorithme utilisant à la fois des données radiologiques et des facteurs de risque connus était supérieur aux simples données radiologiques ou au modèle Tyrer-Cruz utilisé actuellement pour déterminer les femmes à haut risque nécessitant un conseil génétique[26].

2.1.5 Tomosynthèse et dépistage

Le principe de la tomosynthèse (TS) digitale est de pallier à la superposition des structures sur une mammographie 2D grâce à un mouvement du tube à rayon X en arc de cercle et de permettre une meilleure dissociation des structures à la lecture. L'assemblage des coupes par reconstruction permet également la création d'une image dite mammographie synthétisée (MS).

A l'heure actuelle, les recommandations de bonne pratique pour le dépistage du cancer du sein en France et en Europe ne font pas état de la tomosynthèse, chose qui pourrait évoluer en raison d'études récentes montrant une augmentation du taux de détection des cancers dans les populations de dépistage face à la mammographie digitale (MD) seule [27].

Si son utilisation à plus large échelle reste débattue notamment en raison du surcroit d'irradiation qui atteint environ le double de surface mammaire par rapport à une mammographie standard, de nombreux centres en France l'utilisent déjà de manière systématique dans le dépistage du cancer du sein. Mais ces images restent à l'appréciation du premier lecteur et ne sont pas relues en deuxièmes lecture en raison du mode d'envoi physique des clichés.

L'intégration de la tomosynthèse au dépistage passera probablement par le remplacement du couple MD+TS par le couple MS+TS pour optimiser la dose délivrée, même si sa non-infériorité reste à prouver sur de larges études.

Peu de logiciels basés sur le deep learning proposent leurs solutions à la fois sur la mammographie digitale et sur la tomosynthèse, la création d'algorithmes sur une des techniques n'étant pas valable sur l'autre, tout le processus de création et d'entrainement doit être réalisé de façon indépendante. Il est également probable que les Computer-Aided-Detection (CAD) d'imagerie 2D fonctionnant sur les mammographies digitales doivent être réentraîné sur des mammographies synthétisées.

3 Computer-Aided-Detection, start-ups et état de l'art en sénologie

3.1.1 Computer-Aided Detection

Depuis le début des années 2000, des études se sont penchées sur l'intérêt du CAD dans la lecture des mammographies et notamment de la possibilité de remplacer la double lecture par une simple lecture d'un radiologue "augmenté" d'un CAD[28].

Si ces études n'ont pu conclure à une avancée significative en termes de sensibilité de la détection et du taux de rappel, c'est que les CAD dits "traditionnels" étaient peu efficaces.

Ces derniers se basaient sur la détection de la présentation spécifique d'une pathologie sur une image, nécessitant une définition détaillée des caractéristiques de l'image par les développeurs du logiciel[7].

En 1998 aux Etats-Unis, la FDA (Food and Drug Administration) avait approuvée l'utilisation des CAD en mammographie et en 2002 les assureurs ont augmenté le remboursement entraînant une augmentation massive de leur utilisation dans la pratique quotidienne. Près de 80% des radiologues aux Etats-Unis utilisaient alors ces CAD « ancienne génération » en 2015, ce qui explique également l'intérêt prioritaire des start-ups dans ce marché.

Mais une étude de 2015 n'a trouvé aucune amélioration dans la détection des cancers sur une large population avec même une tendance à la baisse de sensibilité, entraînant une augmentation des coûts sans amélioration de la prise en charge des patientes[29].

Les logiciels basés sur le deep learning n'ont pas besoin d'une définition des caractéristiques d'un cancer, le logiciel va trouver lui-même les caractéristiques grâce au labelling du data, sans que l'on sache très bien à quoi elles correspondent.

L'étude de Kooi et al en 2017 a montré que des logiciels sur des réseaux de neurones convolutionnels avaient des performances nettement supérieures au CAD traditionnels[30].

La progression et l'efficacité de ces nouveaux logiciels basés sur l'IA vont permettre de réévaluer le rapport coût-efficacité du dépistage en France qui bénéficie actuellement de la double lecture.

3.1.2 Start-up, développement et modèle économique

Aujourd'hui environs 140 entreprises sont lancées sur le marché de l'IA en radiologie. Dans le secteur de la sénologie, une dizaine de logiciels disposent aujourd'hui d'une FDA clearance, sésame hautement important et nécessaire pour une utilisation sur le

marché américain et vitale pour les différentes levées de fond. Grâce à ces levées de fond, la plupart des entreprises subsistent dans la course aux logiciels, car pour l'instant, les revenus liés à la vente de leurs algorithmes sont faibles voire inexistants. Une grande partie des investissements est placée dans la partie recherche et développement du produit en attendant d'avoir un produit fini et de rediriger les financements vers la promotion et la vente du logiciel. Mais ces derniers devront prouver une amélioration du rapport coût/efficacité et/ou performance pour les centres de radiologie avant d'être implantés.
Le modèle de facturation de ces logiciels est également débattu, la plupart des entreprises s'orientant vers une tarification à l'acte en plus d'un abonnement pour l'accès au logiciel. De plus, du fait de la grande variété des modèles économiques mondiaux et d'organismes de financement privés et publics, il ne pourra pas y avoir un seul type de facturation. Le remboursement par les organismes de santé publique sera un des moteurs de l'implantation des logiciels.

Etant donné la multiplicité des logiciels et l'étroitesse de leur champ d'action, des plateformes mettant en commun une palette de logiciels se sont développés, à l'instar d'Incepto en France ou Arterys aux États-Unis, qui proposent des logiciels dans plusieurs champs d'application radiologique via une plateforme unifiée.

3.1.3 Etats de l'art des logiciels en sénologie

Les logiciels dont la technologie est basée sur l'intelligence artificielle sont majoritairement spécialisés dans un domaine, comme l'analyse de la densité, la détection de lésion en imagerie 2D ou en tomosynthèse. Mais il est probable qu'avec l'évolution du marché, les logiciels proposent peu à peu toutes ces fonctionnalités, que soit par la fusion, le rachat des start-ups ou le développement interne.

3.1.3.1 Évaluation standardisée de la densité mammaire

La densité du sein est classée selon l'ACR (American College of Radiology) en quatre catégories et fait partie du compte rendu systématique de chaque examen. L'évaluation est réalisée par le radiologue en fonction du ratio de glande mammaire par rapport à la graisse contenue dans le sein. Si les catégories extrêmes sont bien départagées par les radiologues, A pour un sein presque entièrement graisseux (<25%

de glande) et D pour un sein complètement dense (>75% de glande) ; les catégories B et C ont une forte part de variabilité inter-observateur et intra-observateur[31]. Plusieurs solutions utilisant le DL proposent une évaluation standardisée et automatique en fonction du pourcentage de glande mammaire sur la mammographie :

-Quantra Breast Density assessment est développé par Hologic pour être à terme intégré à la console du constructeur et possède une approbation FDA depuis 2017. Une étude montre la corrélation face à 3 radiologues[32]. https://www.hologic.com/hologic-products/breast-skeletal/image-analytics

-Volpara Density possède également une approbation FDA depuis plusieurs années, et des études concernant la corrélation du logiciel avec une quantification IRM ou comparée à plusieurs radiologues sont connues depuis 2014[33]. https://www.volparasolutions.com/science-hub/breast-density/

-Densitas intègre directement les données de densité au compte-rendu.

Ces deux dernières solutions proposent également un logiciel de contrôle qualité de l'examen en amont de la visualisation par le radiologue avec monitoring des constantes de la mammographie, et des indicateurs de qualité tout en proposant les solutions pour améliorer l'examen. https://densitas.health/radiology-department

3.1.3.2 CAD & Exam flagging en mammographie standard et tomosynthèse

Des solutions se sont développées, pour le flagging des images, censés optimiser le workflow en mammographie en indiquant directement après la réalisation de l'examen si celui-ci est pathologique ou bénin par un "drapeau". Si la majorité des logiciels proposent un CAD, ces deux technologies nécessitent un entraînement différent de l'algorithme. En effet, le flagging va reposer sur un entraînement semi-supervisé ou seule l'information bénin ou malin sera indiquée dans les annotations liées à l'image, sans que la lésion soit précisée. Le premier concours appelé en 2017 nommé DREAM challenge, qu'une entreprise française (Therapixel) a remporté, reposait sur ce principe. L'entraînement de l'algorithme était réalisé avec un temps et des données limitées. Le fonctionnement du CAD repose lui sur un entraînement complètement supervisé de l'algorithme pour que chaque lésion puisse être spécifiée.

-Mia de Kheiron Medical est une entreprise anglaise qui se positionne comme un deuxième lecteur autonome afin de diminuer les coûts des campagnes de dépistage, et dispose d'une certification européenne (CE). Le logiciel fonctionne de façon

indépendante sur l'ensemble des consoles et permet de donner une décision synthétique globale sur un dossier. Le logiciel dispose ensuite de bounding box afin d'expliquer sa décision. Aucune étude n'est disponible actuellement.

-Therapixel est l'entreprise française ayant remporté le DREAM challenge. Elle travaille aujourd'hui avec un consortium de 13 centres de radiologie en France sous l'appellation MammoScreen[34] dans le but de diminuer les biais liés à des populations, des constructeurs ou des acquisitions différentes selon les pôles d'imagerie. Son logiciel fonctionne avec une console indépendante et un traitement des images via le Cloud.

-Transpara de Screenpoint[35], entraîné initialement sur 9000 mammographies malignes et 9000 bénignes, a prouvé dans une première étude que sur 14 radiologues de niveaux de spécialisation différents, les performances étaient meilleures quand le radiologue était aidé par l'algorithme que lorsqu'il opérait seul. Le logiciel utilisé seul promettait également des performances non inférieures aux radiologues sur les courbes ROC (Receiving Operator Characteristic). L'étude la plus récente comparait les résultats d'études européennes et américaines dans une méta-analyse portant sur 2652 cas, issus de 9 data-sets, entre l'algorithme Transpara et 101 radiologues au total. L'aire sous la courbe de la courbe ROC du logiciel était non-inférieure aux résultats des radiologues d'après les analyses statistiques.

Le logiciel fonctionne avec une délimitation des lésions appelée bounding box où apparaît selon l'algorithme le pourcentage de risque que la lésion soit cancéreuse. Cette fonctionnalité peut être automatique ou à la demande du radiologue lorsqu'il passe son curseur sur la lésion. Le logiciel produit enfin un score de suspicion entre 1 et 100. Il fonctionne également sur l'imagerie en tomosynthèse et possède une clearance FDA[36,37].

-CureMetrix est une entreprise américaine, dont le logiciel cmAssist[38] fonctionne comme une aide au diagnostic avec un CAD, en fournissant un score de suspicion de malignité appelé NeuScore. L'étude a été publiée dans le Journal of Digital Imaging met en évidence une amélioration de la détection de 7,2% dans un set de 122 examens enrichis en cancer, avec 90 cas de faux négatifs, sur un panel de 7 radiologues avec lecture avant et après utilisation du logiciel[39].

-iCAD est spécialisé dans l'imagerie en tomosynthèse avec ProfoundAI[40], dispose d'un CAD classique pour des images en 2D appelé Second Look et fournit un logiciel

d'analyse de la densité. Le logiciel va comme celui de Screenpoint indiquer pour chaque lésion un pourcentage correspondant au risque de malignité selon l'algorithme sur les images en tomosynthèse. Ce logiciel présente une fonctionnalité intéressante qui est une jauge permettant de définir le niveau de sensibilité de l'algorithme selon sa pratique, avec par exemple un niveau de sensibilité abaissé permettant de ne retenir que les lésions ayant le plus de chance d'être malignes. Un score de malignité est également assigné à l'ensemble des études.
Cette solution est compatible avec la plupart des constructeurs, elle fonctionne avec une console séparée, un serveur localisé et sera bientôt intégrée directement à certaines consoles. Une étude en cours de publication montre une augmentation de 5,7% de la performance des radiologues sur les courbes AUC.

-IBM healthcare en partenariat avec le centre Universitaire israélien de Haifa montre dans une étude parue en juin 2019 une amélioration du taux de faux négatifs des radiologues grâce à son algorithme, qui correspondait à 0,8% de la population totale. L'originalité de cette étude est que l'algorithme inclut des données radiologiques et des données clinico-biologiques, et semble s'améliorer quand les deux types de données sont combinées. Mais il existe de nombreux biais dans cette étude où les "bounding box" n'étaient pas présentes pour affirmer que les données sur lesquelles se basait l'algorithme pour la détection des lésions étaient les bonnes, chose d'autant plus importante sur une cohorte comprenant de nombreux faux négatifs des radiologues. De plus l'analyse des facteurs cliniques montre une décorrélation entre le antécédents familiaux et les biopsies positives pour le cancer, ce qui permet d'émettre un doute sur la fiabilité des autres marqueurs dans l'analyse.

-Google Health et sa filiale Deep Mind ont publié une étude en janvier 2020 concernant un algorithme pour le dépistage du cancer du sein sur la plus grosse cohorte jamais étudiée. Cette étude propose 3 différentes analyses, la première avec un data set provenant de la base de données de dépistage du NHS, le système de santé du Royaume-Unis, avec 258 765 mammographies, non enrichies en cancer, où les résultats de l'algorithme ont été comparés aux résultats rétrospectifs de la première lecture, puis aux résultats de la première et seconde lecture combinés. L'algorithme montre une augmentation de la sensibilité et de la spécificité face à la première lecture, et une non-infériorité de ces paramètres face à la seconde lecture, dans un système

qui classe les mammographies de façon dichotomique, en normal ou anormal. Le data set provenant d'un centre américain avec près de 3000 mammographies a été enrichi en lésions cancéreuses, et servait également selon les auteurs à prouver l'utilisabilité de l'algorithme en fonction des populations. Il montre une augmentation de la sensibilité et de la spécificité encore plus forte que dans la cohorte anglaise, ce qui peut s'expliquer par le système de dépistage différent, sans deuxième lecture mais avec une fréquence de dépistage annuelle. Enfin la dernière partie de l'étude confronte 6 lecteurs à un data set restreint de 400 mammographies enrichies, avec une classification basée sur le BI-RADS, dans laquelle les lecteurs avaient également accès au mammographies antérieures. Ici aussi les courbes ROC montrent des performances supérieures de l'algorithme.
Les auteurs suggèrent que leur algorithme pourrait être utilisé pour supprimer la deuxième lecture et ne faire arbitrer par un expert que les cas discordants entre le premier lecteur et l'algorithme.

3.1.3.3 CAD en échographie mammaire

Un système automatisé d'échographie, avec un rendu volumique du sein et un CAD associé sur les images générées, est censé diminuer la variabilité inter-opérateur. Ce type de logiciel développé pour le marché américain a peu de chance de trouver preneur en l'état actuel sur le marché français où les radiologues réalisent eux-mêmes les échographies en fonction des images de mammographie analysées.
-Qview Medical[41] détient une approbation FDA depuis 2016 et permet de diminuer le temps d'analyse sans diminuer la performance de diagnostic.

- Koios Medical[42] : ce logiciel permet l'extraction et l'analyse des caractéristiques d'une zone d'intérêt lorsque le radiologue clique sur une lésion suspecte d'une image échographique disponible dans le PACS.

3.1.3.4 CAD en IRM mammaire

-Merge CAD-stream développé avec Watson-IBM[43] produit un CAD pour IRM avec détection des lésions et optimisation du workflow.

-DynaCAD, de InvivoCorp par Phillips[44], a développé en même temps que son logiciel pour l'IRM de prostate un logiciel de CAD en IRM avec une segmentation automatique des lésions, des courbes de perfusion automatiques avec une organisation et une optimisation du workflow.

-Intelliscan par UK Innovate[45], est le fruit de la collaboration de plusieurs universités anglaises et start-ups pour la création d'une solution permettant la détection automatique des lésions et d'un support de décision dans la prise en charge en IRM mammaire.

MATERIEL ET METHODES

Nous avons réalisé une étude rétrospective, indépendante, de performance, au centre hospitalier de Valenciennes, avec un échantillon test issu de données locales, destiné à évaluer deux logiciels basés sur le deep learning.
De la base de données de Valenciennes, a été tiré un set de data afin de tester initialement le logiciel d'Arterys Breast AI. Il a ensuite été proposé au logiciel de la firme Française Therapixel de tester l'algorithme de façon indépendante avec le même set de données.

Le pôle de radiologie de Valenciennes a eu l'opportunité de développer un partenariat avec Arterys, entreprise américaine, californienne, basée à San Francisco et spécialisée dans le développement de solutions pour la radiologie basées sur le deep learning.

Le prototype de logiciel Breast AI, livré début juin 2019, est un logiciel basé sur le Cloud computing d'Amazon AWS dont l'utilisation se fait grâce au navigateur Chrome de Google. Les mammographies sont téléchargées sur le Cloud, le logiciel fournissant une solution de triage des examens comportant des lésions nécessitant des investigations à réaliser.

La priorité d'Arterys était le développement d'un logiciel de flagging, afin d'optimiser le workflow dans un modèle de dépistage « à l'américaine », l'étude a donc été conçue pour étudier l'aptitude des logiciels à effectuer un tri des mammographies préalable à la lecture. L'objectif de ce type de module étant un gain de temps dans ce modèle de dépistage différé où le radiologue commence sa vacation avec un grand nombre d'examen à interpréter, les études suspectes seraient lues en priorité, et les études négatives pourraient être lues secondairement.

L'étude a ensuite été soumise à un autre logiciel, Mammoscreen de l'entreprise française Thérapixel, dont le logiciel fonctionne tout d'abord comme une aide à la détection, et fournissant un score de suspicion de malignité allant de 1 à 10.

Enfin 3 radiologues ont lu le set de mammographie dans une lecture de type différée et dans les mêmes conditions que l'algorithme.

1 DATA

Les données anonymisées de 140 patientes et de leurs mammographies issues de notre pratique quotidienne, diagnostique et dépistage. Ces études ont été choisies aléatoirement dans la base de données constituée sur le PACS de Valenciennes en fonction du classement Bi-RADS.

Ce set de données a été enrichi en étude classées BI-RADS 3,4,5 par rapport au pourcentage d'étude suspecte en pratique courante dans un centre radiologique.
Les études BI-RADS 3 ont fait l'objet d'une surveillance pendant 2 ans selon les guidelines françaises où ont bénéficié d'une biopsie. Les études BI-RADS 4 et 5 ont bénéficiées d'une étude histologique par biopsie.

La distribution des études est visible sur la table 1 :

Table 1 - Distribution du BI-RADS initial au sein du data-set.

CATEGORY	NR. OF CASES
BI-RADS 1	20/140 (14%)
BI-RADS 2	60/140 (43%)
BI-RADS 3	20/140 (14%)
BI-RADS 4	20/140 (14%)
BI-RADS 5	20/140 (14%)

Au sein de notre échantillon de 140 mammographies, 40 ont été biopsiées, avec 27 lésions malignes et 13 lésions bénignes, comme visible sur la table 2. Sur les 27 lésions malignes, 24 correspondaient à des lésions de carcinome infiltrant, 1 à une

lésion de carcinome in-situ et 2 à des lésions hyperplasie cellulaire atypique, considérées comme positive, car nécessitant un traitement chirurgical.

Table 2 - Distribution des biopsies en function du BI-RADS

CATEGORY	NR. OF BIOPSY	NR. OF BENIGN LESION	NR. OF CANCEROUS LESION
BI-RADS 3	1	1	0
BI-RADS 4	19	10	9
BI-RADS 5	20	2	18

2 OPERATEURS :

2.1 Les algorithmes

-Breast AI de Arterys : L'apprentissage et la validation du modèle ont été effectués à l'aide de 11 244 et 1408 études respectivement. Chaque étude est composée des quatre vues de dépistage standard (L CC, L MLO, R CC, R MLO), sans vues supplémentaires (compression / agrandis). Les images avec implants mammaires ont été exclues.

L'algorithme fonctionnait sur la base d'une architecture ResNet50[46].

La majorité (86%) des données d'apprentissage et de validation proviennent de l'ensemble de données OPTOMAM[47], qui inclut des annotations cliniques et de bounding box. Les données de référence ont été déterminées en prenant le score le plus élevé des annotations cliniques et en convertissant ce score en BIRADS. Des données supplémentaires ont été ajoutées à partir de 3 centres hospitaliers, DASA, St Joseph et Valenciennes, représentant respectivement 4%, 4% et 6%. Pour chacune des sources, les données de référence ont été déterminées à partir du score BI-RADS du rapport du radiologue.

-Mammoscreen de Therapixel :

2.2 Les radiologues

Trois radiologues, deux séniors spécialisés en imagerie de la femme et en sénologie, avec respectivement 11 ans d'expérience pour le lecteur 1 et 14 ans pour le lecteur 2, réalisant tous les 2 des examens en seconde lecture pour le dépistage organisé du

cancer du sein, et un radiologue junior en 9ème semestre de son internat, ont lu les études dans des conditions similaires à celle des algorithmes. Ils disposaient uniquement des vues de face (CC) et oblique (MLO) droites et gauches, sans avoir connaissance des examens antérieurs, des examens complémentaires ou du contexte clinique.

2.3 Analyses

Pour chaque étude, les radiologues ont annoté sur un fichier excel, s'ils considéraient que la mammographie était :
-Négative : sans lésion suspecte identifiée.
-Positive : image nécessitant des examens complémentaires (cliché en compression, agrandis ou échographie).
Ce système de classification tend à s'approcher du fonctionnement du dépistage aux États-Unis et de la notion de "rappel" pour les examens positifs.

Pour des analyses supplémentaires, les radiologues ont aussi annoté dans le fichier, si la nature des lésions était plutôt bénigne ou suspecte, et la latéralité de localisation de la lésion.

Les deux algorithmes commerciaux proposaient des solutions différentes pour la lecture des mammographies, nous avons dû adapter leurs résultats pour effectuer des analyses reproductibles entre les différents acteurs de l'étude.

Le premier algorithme Breast AI d'Arterys était conçu pour "flagger" les études positives nécessitant un rappel.

Le logiciel de Thérapixel, nommé MammoScreen donne un score pour chaque étude entre 1 et 10. Les études disposant d'un score bas ont une faible probabilité d'être maligne pendant que celles qui ont un haut score sont plus susceptibles de contenir une lésion suspecte de malignité.

Nous avons convenu de prendre le score inférieur ou égal à 4 pour distinguer les études négatives des études positives.

Les résultats ont été analysés selon 2 scénarios :

- *Le BI-RADS initial correspond au « ground truth »*

Dans ce scénario, si le BI-RADS initial était 1 ou 2 l'étude était considérée comme négative, et si le BI-RADS contenait une lésion BI-RADS 3,4 ou 5. La définition de la positivité ou de la négativité du test est reportée table 3 :

Table 3 – Examens positifs et négatifs en fonction du BI-RADS

	Initial BI-RADS 1/2	**Initial BI-RADS 3/4/5**
Test positive	False Positive	True positive
Test negative	True Negative	False negative

Pour l'analyse des résultats, nous avons comparé les résultats des radiologues et des différents algorithmes en comparant les performances globales et les performances pour chaque BI-RADS.

- *Le résultat des biopsies correspond au « ground truth »*

Dans cette hypothèse, une lésion cancéreuse prouvée histologiquement correspondait à une étude positive, s'il n'y avait pas eu de biopsie ou si la biopsie révélait une lésion bénigne, l'étude était considérée comme négative. La définition de la positivité ou de la négativité du test est reportée table 4 :

Table 4 – Etudes positives et négatives en fonction du rapport de biopsie.

	Pas de biopsie ou biopsie négative	**Biopsie positive pour lesion maligne**
Test positive	False Positive	True positive
Test negative	True Negative	False negative

Résultats

Neuf études ont dû être exclues du data-set car elles correspondaient à des BI-RADS 3 visibles uniquement en échographie, réalisé pour densité mammaire élevée pendant l'examen.

Trois études BI-RADS 4 étaient visibles uniquement en tomosynthèse, nous avons décidé de laisser ces études dans l'analyse pour tenter de voir si les algorithmes proposaient de meilleures performances.

Les algorithmes et les radiologues avaient des niveaux de performance globale similaires avec un taux de rappel de l'ordre de 50%, qui correspond globalement au taux d'études positives dans notre set de données. Le taux d'examens rappelés par lecteur est visible sur les diagrammes de la figure 1 :

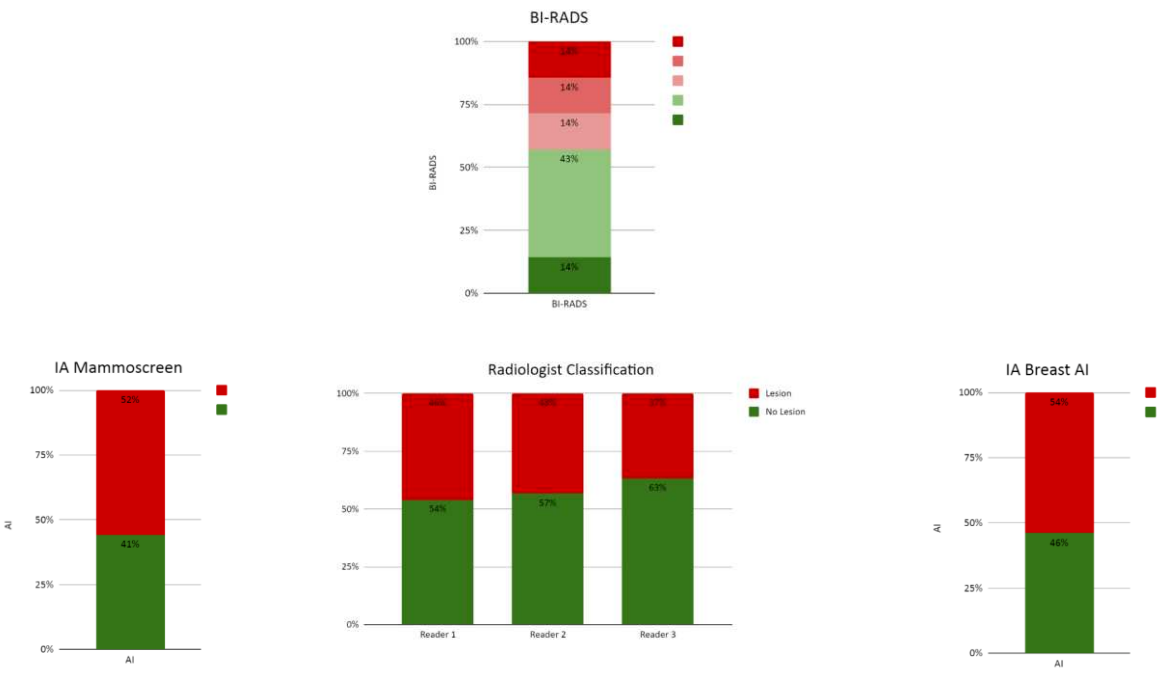


Figure 1 - Taux d'étude positive et négative en fonction des différents lecteurs

L'analyse des performances globales des différents opérateurs pour le « ground truth » dépendant du BI-RADS initial est reportée sur la table 5. Les performances de chaque opérateur et pour chaque type de BI-RADS sont représentées par la figure 2.

Table 5 - Performance des radiologues et des algorithmes en terme de sensibilité (TVP) et du taux de faux positif (TFP) avec la classification BI-RADS initiale comme groudn truth

OPERATOR	TPR	FPR
Algorithm 1 Breast AI	0,78	0,34
Algorithm 2 MammoScreen™	0,71	0,28
Radiologist 1 (R1) Senior	0,82	0,34
Radiologist 2 (R2) Senior	0,76	0,33
Radiologist 3 (R3) Junior	0,71	0,28

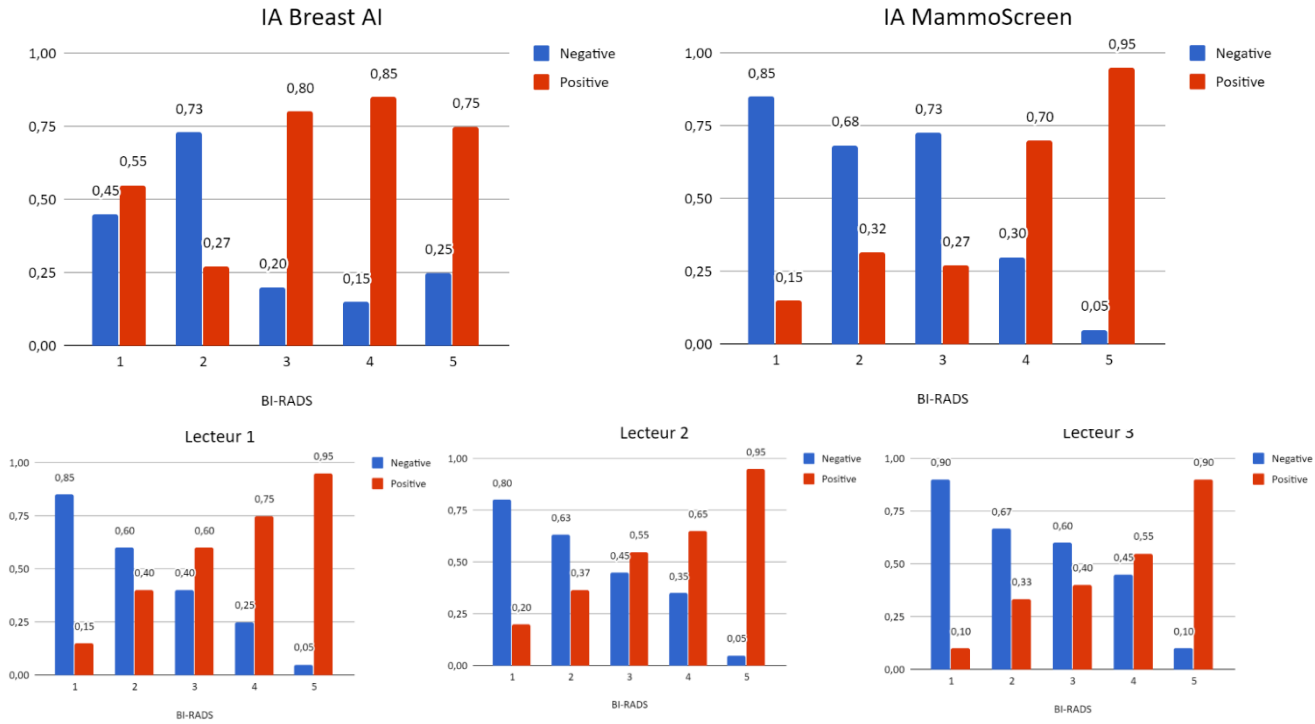


Figure 2- Graphique des analyses par BI-RADS et par opérateur

L'algorithme Breast AI présente des performances globales satisfaisantes, similaires à celles des radiologues experts.Cependant l'analyse par BI-RADS montre des résultats contrastés,l'algorithme ayant classé comme positives 55 % des lésions BI-RADS 1 et n'a pas retenu 25% des lésions BI-RADS 5.

Le logiciel Mammoscreen avait des performances globales moins bonnes que celles des radiologues, mais des performances sur les BI-RADS extrêmes comparables à celle des radiologues experts.

Les performances pour la seconde analyse qui était basée sur le « ground truth » biopsie, sont présentées sur la table 6.

Table 6 - Performances des algorithmes et des radiologues sur le "ground truth" biopsie

OPERATOR	TPR	FPR
Algorithm 1 Breast AI	0,41	0,57
Algorithm 2 MammoScreen™	0,89	0,33
Radiologist 1 (R1) Senior	0,89	0,44
Radiologist 2 (R2) Senior	0,85	0,41
Radiologist 3 (R3) Junior	0,81	0,36

Le premier algorithme montre de mauvaises performances sur le « ground truth » biopsie en termes de sensibilité et du taux de faux positifs. Le second algorithme présente en revanche des performances similaires à celle des radiologues experts en termes de sensibilité et de spécificité.

Les courbes ROC pour les deux types de « ground truth » sont visibles sur la figure 3, avec les AUC (Area Under the Curve) correspondant sur la table 7.

La figure 3 rapporte les courbes ROC (Receiving Operating Caracteristic) pour les deux types de « ground truth ».

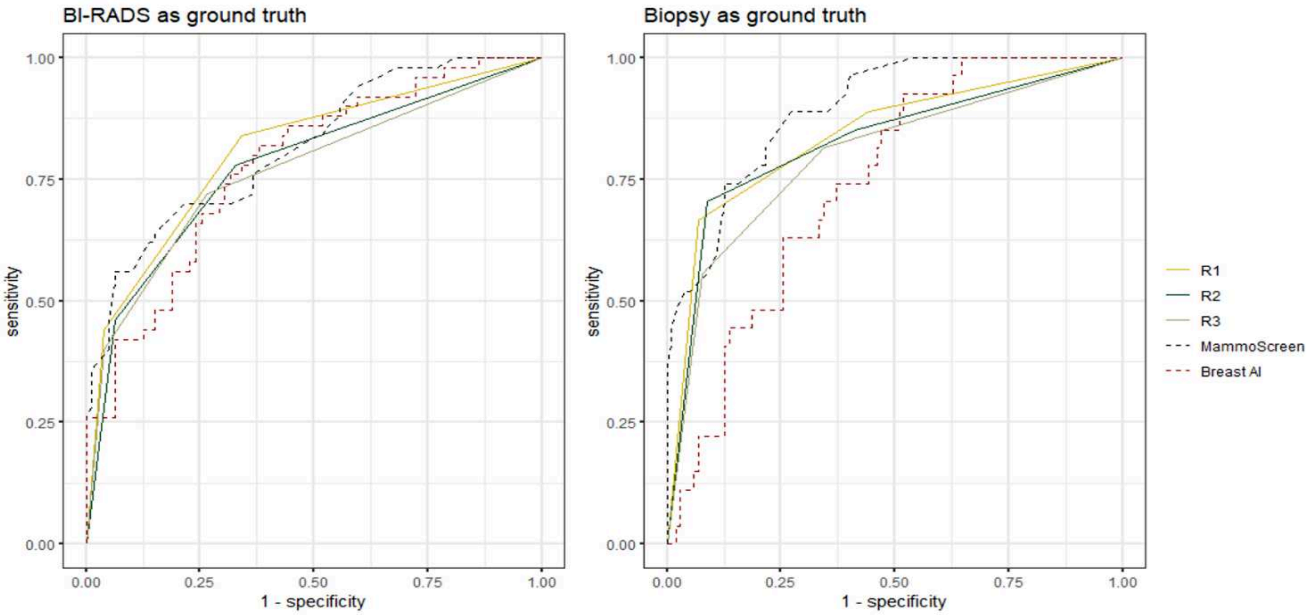


Figure 3 – Courbes ROC utilisant le BI-RADS initial comme ground truth (gauche) et les résultats des biopsies (droite). Les courbes des algorithmes sont construites de façon continue entre 0 et 1 contrairement aux radiologues où seuls trois points étaient disponibles (leurs choix étant 0,1 ou 2).

Table 5 - AUC des opérateurs pour les deux conditions de ground truth.

OPERATOR	AUC (BI-RADS as ground truth)	AUC (biopsy as ground truth)
Algorithm 1 Breast AI	0.778	0.742
Algorithm 2 MammoScreen™	0.814	0.895
Radiologist 1 (R1) Senior	0.808	0.84
Radiologist 2 (R2) Senior	0.777	0.827
Radiologist 3 (R3) Junior	0.767	0.799

Discussion

1 Méthodologie

Lorsque l'étude a été conçue, il s'agissait initialement de tester, dans le cadre du partenariat avec Arterys, son module de triage et les progrès réalisés par l'algorithme. L'idée de challenger des radiologues dans les mêmes conditions de lecture que les logiciels, n'était à postériori par forcément pertinente dans le sens où les conditions de lecture dégradées s'éloignaient de la pratique clinique. Il faut voir les résultats des radiologues comme ceux d'un groupe « contrôle » et nous ne nous attarderons pas sur leurs performances.
Le choix du « ground truth » en fonction du BI-RADS initial nous a permis de décomposer les résultats en 5 catégories et de voir que les performances globales peuvent être mises en défaut sur un set de petite taille.

Le « ground truth » basé sur les biopsies montre quant à lui une disparité de résultats encore plus marquante entre les deux algorithmes, même si le but initial était le triage d'examen nécessitant un « recall », où même des lésions BI-RADS 3, bénignes dans 98% des cas doivent être surveillées. Les performances des radiologues ont donc pu être impactées en terme de spécificité dans le sens où l'exercice auquel ils se sont prêtés leur demandait de retenir toutes les lésions visibles et nécessitant au moins un suivi.

2 Data-set

Notre set de test présentait plusieurs défauts : la population d'étude était une population générale ayant passé des examens à Valenciennes, des examens de dépistage et de diagnostic mêlées, ce qui pouvait induire une difficulté supplémentaire pour les logiciels entraînés sur des Data base majoritairement de dépistage.

Mais en vue de la constitution d'un set idéal en condition réelle d'examen en France, on ne pourrait inclure uniquement des examens de dépistage et l'inclusion des études pour diagnostic est un pas vers une validation externe de qualité des logiciels.

Le choix du nombre de mammographies sélectionnées pour réaliser notre étude est arbitraire. Nous n'avons pas pu nous baser sur une puissance statistique calculée, pour définir le nombre de mammographie adéquat, en raison de l'absence de travaux antérieurs étudiant les logiciels utilisés et d'une estimation précise de leurs niveaux de performance. Nous avons considéré que 140 mammographies était un nombre d'études raisonnable vis à vis des travaux antérieurs concernant ce type de logiciel et de nos impératifs de temps et de moyen.

Le set comprenait des mammographies dont la classification BI-RADS se fixait sur l'échographie réalisée pour densité mammaire élevée durant le même examen, avec des lésions non visibles en mammographie, même après relecture de l'ensemble du dossier sénologique. Ces études ont donc dû être exclues de l'analyse.
Trois études, dont les lésions n'étaient visibles qu'en tomosynthèse ont été maintenues dans l'analyse afin de comparer les performances du logiciel par rapport aux radiologues.

Nous n'avons pas retenu comme facteur confondant le fait que les radiologues travaillaient dans le centre où ont été prélevées les 140 mammographies. En effet, il est possible que les radiologues aient pu déjà avoir interprété ces examens par le passé mais l'échantillon était réparti sur plusieurs années.

Les conditions de lecture pour les radiologues étaient très éloignées des conditions d'exercice classique. En effet, dans l'exercice quotidien le contexte clinique, les examens antérieurs et le Bilan de Diagnostic Initial (BDI) sont capitaux pour la prise en charge des patientes.

Ce type d'étude rétrospective sur des cohortes de mammographies entraîne généralement une baisse de performances des radiologues en comparaison à la pratique clinique, documentée sous le nom de "laboratory effect". Cela pourrait être

expliqué par une baisse de la concentration, sachant que la décision n'a pas d'impact sur les patientes[48].

3 Logiciel Breast AI de Arterys

L'hôpital de Valenciennes et le service d'imagerie de la femme ont proposé leur aide et ont essayé de participer au développement du projet de l'entreprise Arterys et de son logiciel Breast-AI en vue d'une utilisation en pratique clinique. Si les résultats objectivent des résultats bruts avec des performances globales comparables entre le modèle et les radiologues, l'analyse en fonction du Bi-RADS montre que la version actuelle du modèle n'est pas encore utilisable en routine clinique.

En effet, si la performance globale du logiciel, avec une valeur de Taux de Vrai Positif (TVP) à 0,78 et un Taux de Faux Positif (TFP) de 0,34 est satisfaisante et comparable avec les valeurs obtenues par les radiologues seniors, l'objectif principal du logiciel de triage est d'avoir un taux minimal de faux positifs sur les Bi-RADS 1 et un taux de vrai positif maximal sur les Bi-RADS 5 afin d'avoir un gain de temps maximal et se concentrer sur les études complexes.

Dans notre étude, trop de faux positifs rendent inadapté le logiciel à l'élimination des études négatives, et potentiellement chronophage le temps passé à relire des études rendues faussement positives par le logiciel.

Les graphiques de la figure 3 montrent que pour les Bi-RADS 1, le logiciel a arrêté 55% des études et que 25% des cancers Bi-RADS 5 n'ont pas été arrêtés.

En revanche le modèle possède de bonnes performances sur les BI-RADS 2 et 4 seuls.

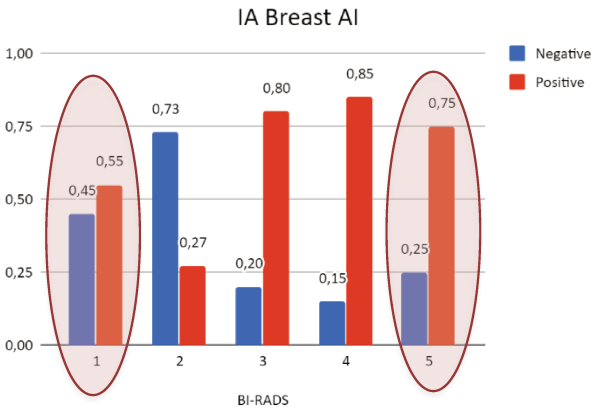


Figure 3 - Répartition par BI-RADS pour le logiciel Breast AI

Les résultats initiaux ont été présentés sous la forme d'une courbe ROC, avec un operating point déterminé par les développeurs afin de maximiser la TVP.

Une fois les résultats décomposés en fonction du BI-RADS, le logiciel s'avérait inutilisable en pratique courante, même en démontrant des performances globales similaires aux radiologues.

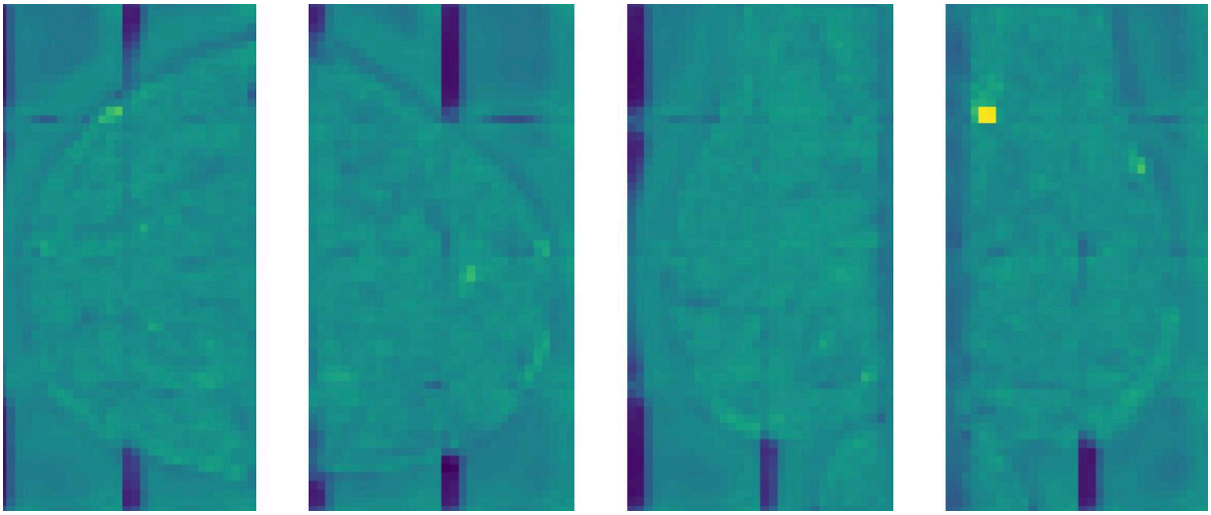
La plupart des études concernant l'intelligence artificielle publiées actuellement présentent leurs résultats sous la forme d'une courbe ROC pour Receiving Operating Characteristic, qui est un outil statistique visuel représentatif et facilement compréhensible. La courbe étant basée sur le rapport sensibilité ou TVP et le TFP ou « 1 – spécificité », les résultats ne sont pas dépendants de la prévalence de la maladie. La performance d'un modèle est établie pour tous ses seuils de performance, définis par les couples TVP (sensibilité) et TFP.
Elles sont utilisées depuis de nombreuses années dans le domaine médical et notamment en radiologie, et permettent d'établir une relation de corrélation entre la sensibilité et le taux de faux positif d'un test diagnostic, calculée pour les valeurs seuils du test[49].
L'analyse des courbes ROC des algorithmes dans notre étude peut être potentiellement confusiogène pour le premier type de « ground truth », celles-ci se croisant à plusieurs reprises et il convient de les analyser avec précaution en prenant en compte l'operating point choisis.

Il faut également être vigilant lors de la présentation des résultats à ce que les courbes et donc les AUC soient comparables, bâties de façon similaire pour chaque opérateur. En effet sans lissage de la courbe, plus le nombre de points constituant celle-ci est faible plus l'AUC va être diminué, induisant un biais en défaveur des opérateurs humains.
Il est possible également d'utiliser les courbes ROC pour définir le seuil optimal d'un test, mais uniquement en cas de valeur continue des données, ce qui n'est pas le cas dans notre analyse.

Les courbes PRC (Precision-Recall Curves) seraient également un outil particulièrement utile sur des sets de données déséquilibrés, avec une forte proportion d'étude négative[50], mais ne peuvent être utilisées sur des sets de données modifiées avec un enrichissement en cancer car ces courbes sont dépendantes de la Valeur Prédictive Positive (VPP), elle-même dépendante de la prévalence de la maladie.

L'analyse des résultats pour le « ground truth » biopsie, conforte les résultats visibles au préalable, avec une sensibilité inférieure à 50%, correspondant à un test non discriminant, et un taux de faux positif trop important.

Pour comprendre la cause des faux positifs et faux négatifs du modèle, certaines des Class Activation Maps (CAMPS) ont été fournies, ces cartes correspondent aux zones activées par le logiciel pour la reconnaissance des lésions.



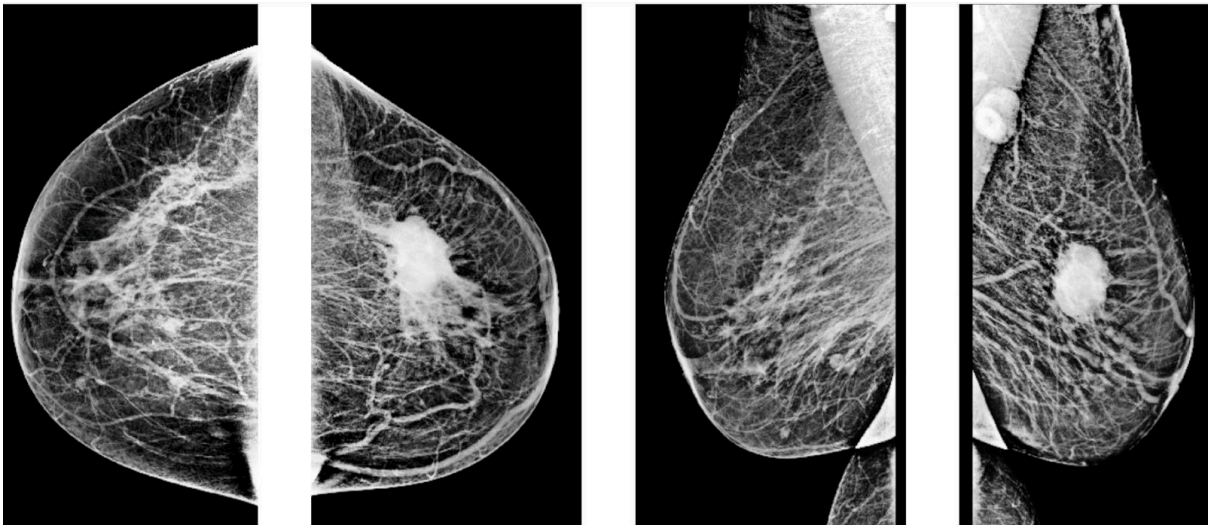


Figure 4 - Faux négatif du modèle avec CAMPs et image de la mammographie correspondante

L'image ci-contre correspond à un faux négatif du modèle pour une masse de 5 cm. La carte d'activation montre que la masse n'a pas été retenue malgré son caractère évident pour un radiologue.

L'absence de masse de grande taille dans la population du set d'entraînement a entraîné la gestion de cette lésion par l'algorithme comme une donnée aberrante ou "outlier" et a conduit l'algorithme à ne pas retenir la lésion.

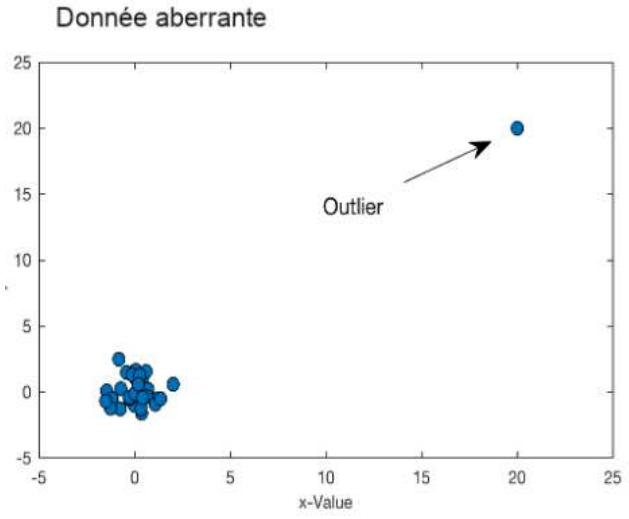


Figure 5 - schéma représentatif de données aberrantes

Une donnée aberrante est une donnée extrême, anormalement différente de la valeur de distribution d'une variable dans la population d'étude[51].

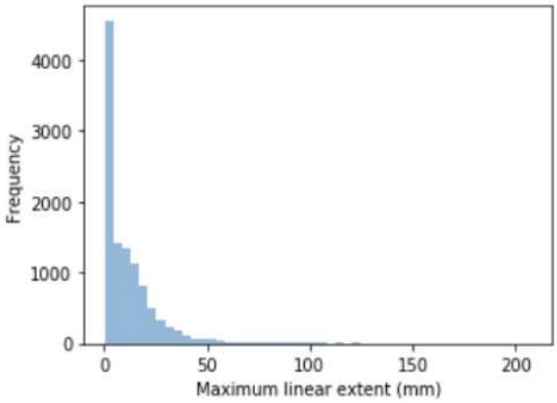


Figure 6 - Répartition dans les données d'entrainement des lésions en fonction de leur taille.

La Figure 6 illustre par un graphique, la fréquence en fonction de la taille des lésions présentes sur les mammographies ayant servies à entraîner l'algorithme. On peut voir que l'algorithme Breast AI a été entraîné sur très peu de donnée contenant des lésions de grande taille, avec une très faible proportion de lésion dont la taille dépasse 5 cm.

Un des arguments avancés par l'entreprise pour expliquer les mauvaises performances de son logiciel était que le data-set utilisé pour l'entraînement de l'algorithme contenait uniquement des mammographies de dépistage, quand notre data-set contenait à la fois des examens de dépistage et des examens diagnostiques. Nous avons donc effectué une analyse en sous-groupe en sélectionnant uniquement les études qui comportaient le mot "dépistage" dans l'indication et nous avons appliqué les mêmes tests de performances, 76 études ont été incluses.
Les performances globales du logiciel ont en effet augmenté, avec une TVP à 0,85 et une TFP à 0,32. Mais l'analyse par BI-RADS met en évidence que si le modèle a très bien répondu sur les BI-RADS 3 et 4, le modèle a arrêté 67% des examens Bi-RADS 1 et a un taux de faux négatifs de 40% sur les Bi-RADS 5, comme visible sur la figure 7.

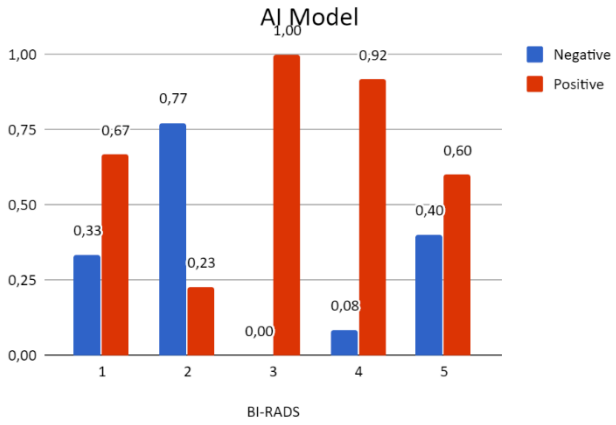


Figure 7 - Répartition par BI-RADS pour le logiciel Breast AI uniquement sur les données contenant le mot « dépistage » dans leur indication.

La population d'étude différente de la population d'entraînement ne peut à elle seule être responsable des mauvaises performances du logiciel sur les Bi-RADS extrêmes, enjeu capital pour l'utilisation du logiciel en pratique courante.

4 Logiciel MammoScreen de Therapixel

Plusieurs scénarios ont été envisagés afin de maximiser le TVP et le TFP en fonction du seuil de score à prendre en compte et c'est le score inférieur ou égal à 4 pour une étude négative et supérieur ou égal à 5 pour une étude positive qui a été retenu en accord avec leur équipe de recherche.

Selon le premier « ground truth », les performances globales de l'algorithme Mammoscreen sont de 0,71 pour le TVP ou sensibilité et de 0,28 pour le TFP. La sensibilité est légèrement inférieure au premier algorithme et est similaire à la sensibilité du radiologue junior, à 0,71. Le TFP est lui légèrement meilleur que celui des autres opérateurs.
La répartition par Bi-RADS montre une corrélation avec les radiologues seniors notamment sur les Bi-RADS 1, 2 et 4, 5.

La perte de performance en TVP se fait donc sur les Bi-RADS 3 pour lesquels le TVP est de 0,27. La répartition des scores sur les Bi-RADS 3 montre 5 cas classés avec le score 4 et 3 avec le score 3. La répartition des scores est visible sur la figure 8 :

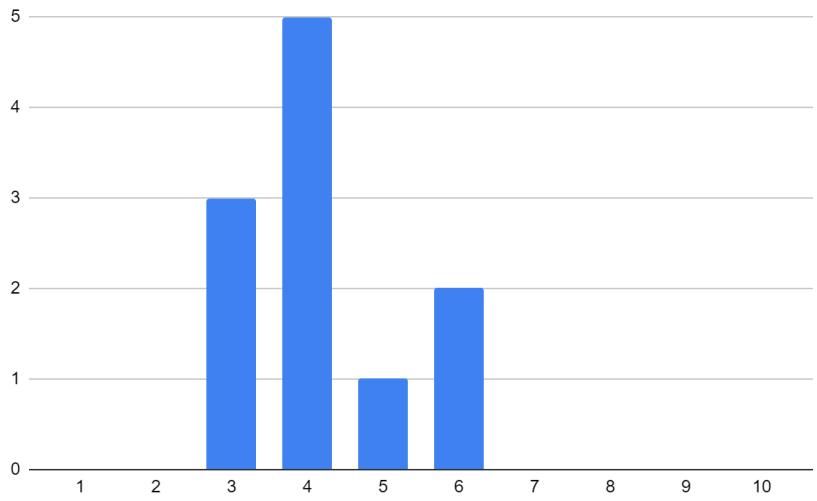


Figure 8 - Graphique de la répartition des BI-RADS 3 en fonction du score du logiciel MammoScreen.

L'analyse de l'ensemble des cas inclus dans le set montre que les Bi-RADS 3, après suivi ou biopsie ne comprenaient que des lésions bénignes, le fait que le score de MammoScreen soit d'abord un score de malignité avant d'être un score pour le triage, a pu jouer en sa défaveur.

A noter que sur l'ensemble des études Bi-RADS 4 et 5, aucune n'a obtenu le score maximal de 10, ce qui correspond selon l'entreprise Thérapixel a un défaut de calibrage du seuil maximal.

L'étude des cas se révèle intéressante car elle met en lumière plusieurs mécanismes de fonctionnement de l'algorithme.
Comme pour Arterys, le logiciel est passé à côté d'une masse très volumineuse, comme visible sur la figure 9 :

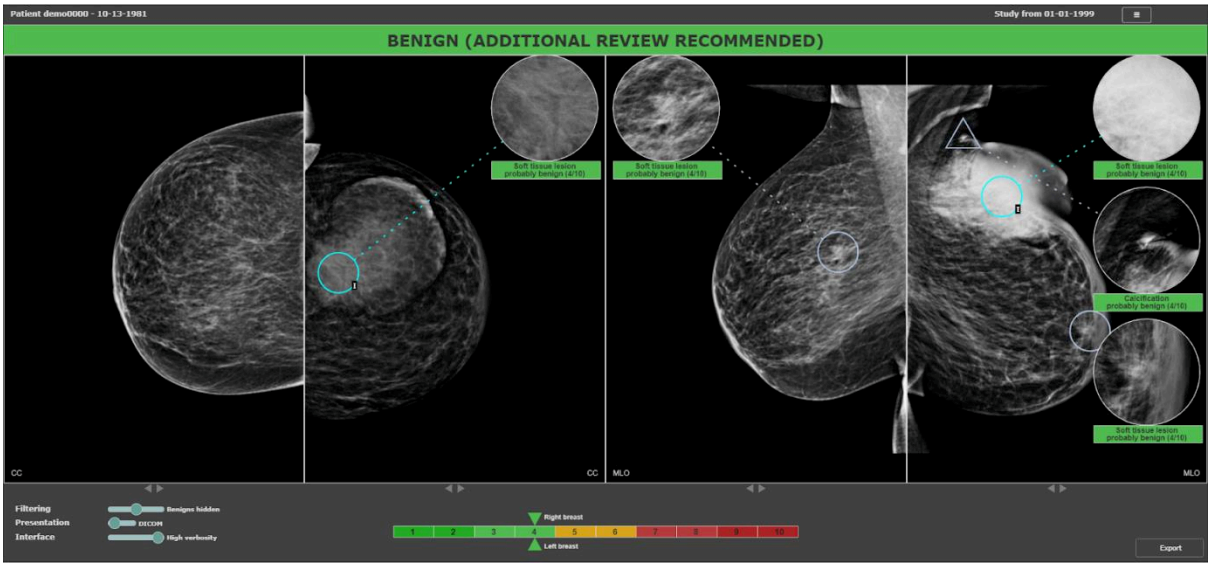


Figure 9 - Capture d'écran du logiciel MammoScreen, ne montrant pas de lésion suspecte.

Plusieurs algorithmes travaillent conjointement dans ce produit, le premier va détecter la masse, mais ensuite l'image va être découpée en patch, qui représente ici un carré de densité élevée et homogène en raison de la grande taille, sans les contours déterminants pour la reconnaissance d'une lésion pathologique.
Si la détection de la lésion tumorale est évidente pour n'importe quel opérateur humain, ce cas illustre encore une fois l'importance des données d'entraînement, ou peu de lésions de grande taille sont présentes dans les populations de dépistage. En pratique, l'utilisation de l'algorithme se fait pour ce logiciel conjointement à la lecture radiologique, le radiologue doit rester attentif et ne requérir l'aide du logiciel que pour les cas litigieux.

Le cas présenté sur la figure 10 ci-dessous est issus cette fois-ci des faux positifs du logiciel MammoScreen, la lésion détectée correspond à une lésion de cytostéatonécrose séquellaire d'une chirurgie. Les données d'entraînement étant dépourvues de patiente opérée, ou d'images post-opératoires comportant des clips ou cicatrices, l'algorithme va être confronté à des données inconnues et prises à tort pour des lésions suspectes.

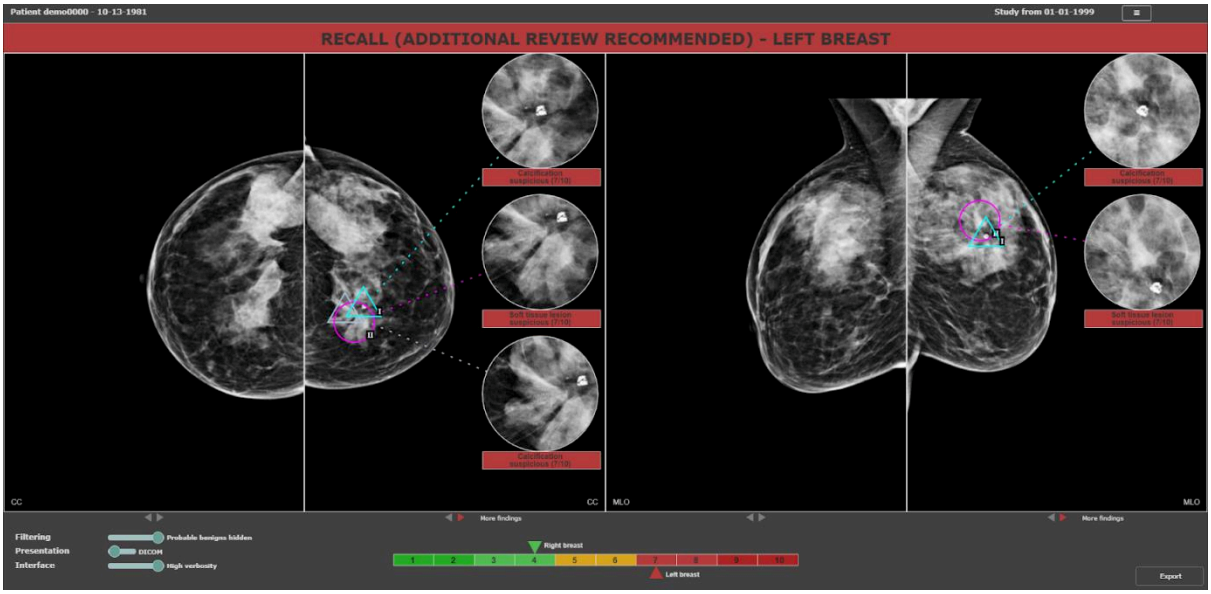


Figure 10 - Capture d'écran du logiciel MammoScreen, montrant une lésion suspecte à gauche.

Un des biais de notre étude provient du fait que nous n'avons pas pu contrôler pour l'ensemble des opérateurs si le classement des études positives et négatives provenait de la détection de la bonne lésion. Pour Arterys, seules certaines des CAMPs ont été fournies, et les radiologues ne disposaient pas d'une grille d'annotation qui détaillait explicitement quelle lésion était détectée.

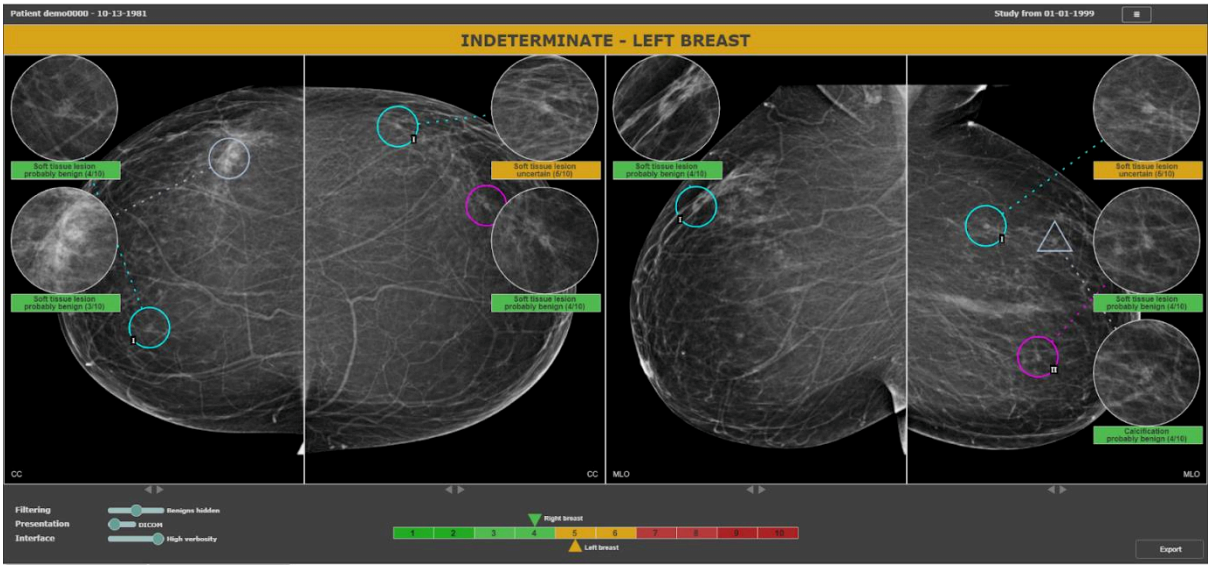


Figure 11 - Capture d'écran du logiciel MammoScreen, montrant une lésion suspecte à gauche, ne correspondant pas après relecture du dossier à la lésion biopsiée.

Le cas ci-dessus, issu des données de MammoScreen, est un cas discordant entre l'algorithme et les radiologues. Deux radiologues sur trois n'ont pas retenu l'étude comme pathologique contrairement à l'algorithme.
L'algorithme a détecté l'examen comme pathologique, avec un score de 5/10.
La lésion pour laquelle le dossier a été arrêté, après relecture du dossier, des comptes rendus radiologiques et anatomopathologique ne correspond pas à la bonne lésion qui a par la suite été biopsiée. Celle-ci correspondait au triangle gris sur la capture d'écran du logiciel, classé comme bénin par l'algorithme avec un score de 4/10.
Mais nous n'avons pu déterminer si le radiologue ayant classé l'étude comme positive avait retenu la bonne lésion, il faudrait donc que les radiologues disposent également d'un outil d'annotation pour la réalisation des prochaines études.

5 Perspectives

Lors de notre étude, l'analyse des examens était très éloignée de la pratique clinique française : aucune information disponible concernant le contexte clinique ou l'indication de l'examen. Les clichés antérieurs et les examens complémentaire n'étaient pas disponibles.
Si cette pratique tend à s'approcher du dépistage organisé aux États-Unis, le fonctionnement du système de dépistage en France fait qu'un système de triage uniquement n'a que peu d'intérêt.
Mais cette étude nous permet d'avoir un premier aperçu de la structure, des biais et limites dont il faudra s'affranchir avant de pouvoir effectuer une validation robuste d'un logiciel.

Pour se rapprocher au maximum de la pratique médicale quotidienne, dans l'hypothèse d'une étude prospective, le nombre d'examens analysés devrait être extrêmement grand. Cela s'explique par le faible taux de cancer dans la population féminine mondiale de dépistage[52] et la nécessité d'un recul de deux ans pour être sûr qu'un cancer de l'intervalle ne soit pas apparu entre deux mammographies de dépistage. La mise en place d'une étude prospective robuste ne pourrait se faire que

sur un temps long, avec une analyse multicentrique et d'importantes ressources engagées.
La validation des algorithmes sur les appareils radiologiques des différentes marques de constructeurs représentés sur le marché sera également nécessaire.
En attendant, il est possible d'effectuer des études rétrospectives robustes avec des échantillons de plus en plus grands afin d'avoir une validité externe des tests de logiciels satisfaisante.
L'étude compilant les résultats du DM DREAM Challenge, qui a eu lieu entre 2016 et 2017 réunissant 125 équipes de chercheurs, industriels et start-up, dont le gagnant fut l'équipe de Therapixel, suggère que la combinaison des algorithmes et des radiologues est supérieure à celle des opérateurs seuls dans un cadre de lecture unique[53].

Actuellement la priorité des start-ups est de développer un produit avec une « FDA clearance », autrement-dit une autorisation d'utilisation et de vente sur le marché américain, qui historiquement dispose d'une force de vente conséquente et des systèmes de cotation et de remboursement des CAD traditionnels mis en place depuis plusieurs années. Pour obtenir une "FDA Clearance", les entreprises doivent envoyer une "premarket notification submission" ou 510(k) à la FDA, qui s'occupera ensuite de comparer l'efficacité et l'innocuité du dispositif avec d'autre logiciels similaires. Ce point est capital pour les investisseurs et nous avons vu que la grande majorité des start-ups développant des systèmes basés sur l'IA fonctionnent grâce à des levées de fond. La lecture décentralisée et différée à partir de quatre vues (CC/MLO) est parfaitement adaptée à des lectures aidées par des logiciels. Leur développement est donc axé sur la pratique américaine avec un marché conséquent du dépistage du cancer du sein, les mammographies se faisant annuellement à partir de 40 ans pour la majorité des femmes.
Pour la certification, les développeurs vont présenter les bases de données sur lesquelles ont été entraînés les algorithmes, ces derniers ne pourront être modifiés à posteriori sans devoir passer par une nouvelle certification.

Concernant la facturation des actes et le paiement des start-ups, ces dernières aimeraient opter majoritairement pour un forfait de licences d'utilisation du logiciel associé à un prix d'utilisation à l'acte. Dans le cadre du dépistage, le prix de revient

pour une mammographie en France, codé par l'acte QEQK004 est de 66,42 euros. Une fois déduit le coût machine, du personnel, le montant restant pour la rémunération oscille à 40% du prix initial, soit aux alentours de 26 euros. Les radiologues pourront-ils assurer le prix de ces logiciels sur leur rémunération sans prise en charge ou revalorisation du prix de l'acte ?

Pour s'assurer d'un fonctionnement économique pérenne sans toucher aux revenus des radiologues, les logiciels étiquetés IA vont devoir prouver leur intérêt sur des études à grande échelle avant d'espérer obtenir une part de remboursement qui grossirait le budget du dépistage.
Si leur supériorité ou leur non-infériorité est prouvée par rapport à la seconde lecture, les instances de remboursement pourraient alors repenser le système de double lecture en la transformant une double lecture radiologue/IA ou une lecture unique mais supervisée par l'IA. Cette réorganisation pourrait alors redistribuer le budget ou tout du moins une partie vers un remboursement de ces algorithmes.
C'est en tout cas le postulat de l'équipe de recherche de Deep Mind dirigée par *McKinney et al,* qui envisage, via l'algorithme de Google dans le modèle de dépistage du NHSS du Royaume-Unis, une refonte du système avec une deuxième lecture algorithmique et un arbitrage expert en cas de discordance homme-machine[54]. Mais les résultats très performants, avec une supériorité de l'algorithme sur la première lecture et une non-infériorité avec la double lecture combinée en termes de mammographie contenant des lésions cancéreuses ne doivent pas faire oublier que dans la pratique clinique, le radiologue doit décider de banaliser ou d'effectuer une surveillance pour nombre de lésion d'allure bénigne. Il serait simpliste de réduire le dépistage à l'unique détection des lésions cancéreuses.
Il conviendra donc de se prononcer sur le type d'algorithme voulu par la communauté et les instances décisionnelles de santé publique en France, une simple lecture aidée d'un algorithme qui aurait des performances non inférieures à la double lecture dans le dépistage ou un véritable remplacement de la deuxième lecture par un algorithme seul. Si cette dernière approche séduit le milieu des investisseurs, un avis d'expert sera encore nécessaire en cas de contradiction entre le radiologue et son algorithme et il nous semble qu'en France, une lecture aidée supervisée serait plus en adéquation avec nos pratiques.

Se poseront également des questions sur la mise en place de ces logiciels, l'application dans les différentes entités de la radiologie en France, à savoir les CHU, les CHn ainsi que les cabinets privés, afin d'avoir une harmonisation des pratiques à l'échelle du territoire français.

L'intégration des logiciels d'IA directement au sein des PACS et des consoles de lecture constituera la prochaine étape dans l'optimisation du workflow et sera absolument nécessaire à l'intégration des solutions dans notre pratique quotidienne[55].

Nous devons signaler que dans les suites de notre étude, la start-up Arterys a changé sa stratégie commerciale et abandonné le développement de son produit Breast AI pour se recentrer sur son activité princeps, l'imagerie cardio-thoracique et le développement d'une plate-forme de streaming pour logiciels.
Lors du partenariat, nous avons eu l'occasion d'échanger avec l'équipe chargée de la recherche et développement du produit (R&D), lors de plusieurs visioconférences entre les États-Unis, le Canada et la France. Si ces échanges ont été enrichissant, il est apparu un manque de compréhension réciproque entre un monde très technique d'une part, et nos réalités cliniques quotidiennes.
S'il manque indéniablement des bases techniques aux radiologues pour arriver à un langage commun, ceux qui développent les logiciels doivent se concentrer sur la pratique clinique des praticiens en s'intéressant à leurs méthodes de travail, au déroulement des vacations et aux problèmes quotidiens pour l'amélioration de leur outil. Si le développement économique du produit se fait grâce aux investisseurs, il ne faut pas perdre de vue que c'est le radiologue qui l'utilisera quotidiennement.

Le second logiciel que nous avons testé de l'équipe Française Thérapixel a d'emblée intégré des radiologues au processus de développement du logiciel. Le logiciel semble plus adapté également à une pratique Française et Européenne, avec un logiciel utilisable comme un assistant de lecture aidant à confirmer ou infirmer des hypothèses diagnostiques lors d'une lecture en temps réel.

Si de nombreuses entreprises sont actuellement en course sur des produits similaires connotés IA, seul un petit nombre présentera dans les années à venir des solutions efficaces et pérennes économiquement. Ces entreprises seront celles qui seront

plébiscitées par les centres de radiologie mais également soutenues par l'industrie et les investisseurs, le retour sur investissement s'avère en effet plus long que prévu, et seules les start-up disposant de fonds économiques suffisant pourront encore être sur le marché dans quelques années[56].

L'intégration des examens antérieurs de suivi dans l'analyse des mammographies lors du dépistage sera une étape déterminante dans l'évolution des logiciels et de leurs performances. Mais cette implémentation n'est pas aisée pour les constructeurs avec comme principal problème le manque de données comprenant des antériorités disponibles pour l'entraînement.

L'équipe de Valenciennes prépare actuellement la réalisation d'une seconde étude dédiée au logiciel Mammoscreen de Therapixel, en essayant de corriger les biais identifiés dans la première étude.

Conclusion

Avant leur implantation à grande échelle, les logiciels d'IA vont devoir passer l'épreuve d'études robustes, multicentriques et à grande échelle, en comparant leurs performances avec les radiologues dans les conditions habituelles d'exercice et disposant de toutes les ressources disponibles d'un centre de radiologie. L'amélioration notable des performances des logiciels passera, notamment dans le dépistage, par l'intégration des examens antérieurs dans les algorithmes.

De façon globale, les logiciels se concentrent actuellement sur la détection des lésions malignes, trouver plus de lésions conduira à plus de traitements, et il faudra prouver une amélioration globale sur la morbi-mortalité de la population. Il est possible que le véritable enjeu de l'IA en sénologie concerne une meilleure sélection et orientation des lésions détectées pour leur prise en charge ou leur surveillance.

Le logiciel d'IA idéal n'existe pas encore, et fonction de l'évolution de l'organisation du dépistage, celui-ci devra fournir une palette d'outil au sénologue. Cette boite à outil comprendra, en amont de la réalisation de l'examen des outils d'adaptation de dose, de la compression avec une optimisation des paramètres techniques et du workflow pour le radiologue et le manipulateur. Le logiciel apportera une réponse adaptée à chaque patiente avec le calcul d'un score de risque en fonction de paramètres cliniques, génétiques, biologiques et radiologiques. Grâce à ce score, les patientes bénéficieront d'une prédiction personnalisée de l'intervalle entre chaque mammographie et la modalité d'examen, la tomosynthèse étant amenée à prendre une part importante dans le dépistage. Enfin lors de l'examen lui-même, le logiciel aidera le radiologue en améliorant sa spécificité et sa sensibilité, l'aidant à réduire le taux de faux positif et de faux négatif, tout en récupérant des informations de façon automatisée sur la densité.
Avec une intégration adaptée à l'environnement des radiologues et à leurs consoles de travail, une confiance accrue dans les diagnostiques des algorithmes, et une amélioration globale du rapport coût-efficacité, le quotidien des praticiens est susceptible de changer dans un futur proche.

En attendant ces évolutions, les radiologues devraient mutualiser et standardiser leurs données afin de constituer des bases de données robustes, de valoriser leurs data et tester de façon indépendante les logiciels développés par les industriels.

Liste des tables

Liste des figures

[Cette liste se met à jour automatiquement. Supprimez ce message.]

Références

[1] LeCun, Y., Bengio, Y. and Hinton, G., “Deep learning,” Nature **521**(7553), 436–444 (2015).

[2] Erickson, B. J., Korfiatis, P., Akkus, Z. and Kline, T. L., “Machine Learning for Medical Imaging,” RadioGraphics **37**(2), 505–515 (2017).

[3] Topol, E. J., “High-performance medicine: the convergence of human and artificial intelligence,” Nat. Med. **25**(1), 44–56 (2019).

[4] Kobayashi, Y., Ishibashi, M. and Kobayashi, H., “How will ‘democratization of artificial intelligence’ change the future of radiologists?,” Jpn. J. Radiol. **37**(1), 9–14 (2019).

[5] Waymel, Q., “Impact of the rise of artificial intelligence in radiology: What do radiologists think?,” 10.

[6] England, J. R. and Cheng, P. M., “Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers,” Am. J. Roentgenol., 1–7 (2018).

[7] Eghtedari, M., “Hallway Conversations in Physics: *What Should a Radiologist Know About Artificial Intelligence?*,” Am. J. Roentgenol. **211**(6), W298–W300 (2018).

[8] “www.lemonde.fr/sciences/article/2019/02/25/les-bugs-de-l-intelligence-artificielle_5428168_1650684.html.”, .

[9] Banks, M. A., “Sizing up big data,” Nat. Med. **26**(1), 5–6 (2020).

[10] “www.sfrnet.org/Data/upload/Images/COM%25/Voyage%20au%20c%C5%93ur%20des%20r%C3%A9seaux%20d%E2%80%99imagerie%20m%C3%A9dicale.pdf.”, .

[11] “Présentation JFR 2019 JP-Beregi www.jfr.radiologie.fr.”, .

[12] “Ethical Dimensions of Using Artificial Intelligence in Health Care.”, AMA J. Ethics **21**(2), E121-124 (2019).

[13] “Making Policy on Augmented Intelligence in Health Care.”, AMA J. Ethics **21**(2), E188-191 (2019).

[14] “moralmachine.mit.edu/hl/fr.”, .

[15] Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Borondy Kitts, A., Birch, J., Shields, W. F., van den Hoven van Genderen, R., Kotter, E., Wawira Gichoya, J., Cook, T. S., Morgan, M. B., Tang, A., Safdar, N. M. and Kohli, M., “Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement,” Radiology **293**(2), 436–440 (2019).

[16] “www.indsante.fr/sites/default/files/Documents_publics/Mode_demploi_-_MR-004_0_0.pdf.”, .

[17] “www.cnil.fr/sites/default/files/atoms/files/mr-003.pdf.”, .

[18] "www.auntminnie.com/index.aspx?sec=sup&sub=aic&pag=dis&ItemID=121679&wf=7612.", .

[19] Finlayson, S. G., Lee, H., Kohane, I. S. and Oakden-Rayner, L., "Towards generative adversarial networks as a new paradigm for radiology education," ArXiv181201547 Cs (2018).

[20] Beers, A., Brown, J., Chang, K., Campbell, J. P., Ostmo, S., Chiang, M. F. and Kalpathy-Cramer, J., "High-resolution medical image synthesis using progressively grown generative adversarial networks," ArXiv180503144 Cs (2018).

[21] "www.has-sante.fr/upload/docs/application/pdf/2013-10/note_de_cadrage_-_depistage_du_cancer_du_sein_chez_les_femmes_de_40-49_ans_et_70-79_ans.pdf.", .

[22] "www.has-sante.fr/portail/jcms/r_1501534/fr/depistage-du-cancer-du-sein.", .

[23] "www.castlighthealth.com/costliest-cities-2015/.", .

[24] Williams, J., Garvican, L., Tosteson, A. N. A., Goodman, D. C. and Onega, T., "Breast cancer screening in England and the United States: a comparison of provision and utilisation," Int. J. Public Health **60**(8), 881–890 (2015).

[25] Pace, L. E., He, Y. and Keating, N. L., "Trends in mammography screening rates after publication of the 2009 US Preventive Services Task Force recommendations: Mammography After the USPSTF Guidelines," Cancer **119**(14), 2518–2523 (2013).

[26] Yala, A., Lehman, C., Schuster, T., Portnoi, T. and Barzilay, R., "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction," Radiology **292**(1), 60–66 (2019).

[27] Thomassin-Naggara, I., Perrot, N., Dechoux, S., Ribeiro, C., Chopier, J. and de Bazelaire, C., "Added value of one-view breast tomosynthesis combined with digital mammography according to reader experience," Eur. J. Radiol. **84**(2), 235–241 (2015).

[28] Gromet, M., "Comparison of Computer-Aided Detection to Double Reading of Screening Mammograms: Review of 231,221 Mammograms," Am. J. Roentgenol. **190**(4), 854–859 (2008).

[29] Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A. and Miglioretti, D. L., "Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection," JAMA Intern. Med. **175**(11), 1828 (2015).

[30] Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A. and Karssemeijer, N., "Large scale deep learning for computer aided detection of mammographic lesions," Med. Image Anal. **35**, 303–312 (2017).

[31] Ooms, E. A., Zonderland, H. M., Eijkemans, M. J. C., Kriege, M., Mahdavian Delavary, B., Burger, C. W. and Ansink, A. C., "Mammography: Interobserver variability in breast density assessment," The Breast **16**(6), 568–576 (2007).

[32] Pahwa, S., Hari, S., Thulkar, S. and Angraal, S., “Evaluation of breast parenchymal density with QUANTRA software,” Indian J. Radiol. Imaging **25**(4), 391 (2015).

[33] Lee, H. N., Sohn, Y.-M. and Han, K. H., “Comparison of mammographic density estimation by Volpara software with radiologists’ visual assessment: analysis of clinical–radiologic factors affecting discrepancy between them,” Acta Radiol. **56**(9), 1061–1068 (2015).

[34] “www.mammoscreen.com/.”, .

[35] “www.screenpoint-medical.com/transpara.”, .

[36] Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Tan, T., Mertelmeier, T., Wallis, M. G., Andersson, I., Zackrisson, S., Mann, R. M. and Sechopoulos, I., “Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists,” JNCI J. Natl. Cancer Inst. **111**(9), 916–922 (2019).

[37] Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.-J., Schilling, K., Heywang-Köbrunner, S. H., Sechopoulos, I. and Mann, R. M., “Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System,” Radiology **290**(2), 305–314 (2019).

[38] “www.curemetrix.com/our-products/cm-assist/.”, .

[39] Watanabe, A. T., Lim, V., Vu, H. X., Chim, R., Weise, E., Liu, J., Bradley, W. G. and Comstock, C. E., “Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography,” J. Digit. Imaging **32**(4), 625–637 (2019).

[40] “www.icadmed.com/assets/dmm252_profound_ai_for_breast_tomosynthesis_revb.pdf.”, .

[41] “www.qviewmedical.com.”, .

[42] “www.koiosmedical.com/solutions.”, .

[43] “www.ibm.com/us-en/marketplace/merge-cadstream.”, .

[44] “www.invivocorp.com/solutions/.”, .

[45] “www.brunel.ac.uk/research/Projects/Enhanced-Artificial-Intelligence-Breast-MRI-Scanning-System-IntelliScan.”, .

[46] He, K., Zhang, X., Ren, S. and Sun, J., “Deep Residual Learning for Image Recognition,” ArXiv151203385 Cs (2015).

[47] “www.cdn.ymaws.com/siim.org/resource/resmgr/siim2017/abstracts/analytics3-Patel.pdf.”, .

[48] Gur, D., Bandos, A. I., Cohen, C. S., Hakim, C. M., Hardesty, L. A., Ganott, M. A., Perrin, R. L., Poller, W. R., Shah, R., Sumkin, J. H., Wallace, L. P. and Rockette, H. E., “The ‘Laboratory’ Effect: Comparing Radiologists’ Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations,” Radiology **249**(1), 47–53 (2008).

[49] van Erkel, A. R. and Pattynama, P. M. T., “Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology,” Eur. J. Radiol. **27**(2), 88–94 (1998).

[50] Saito, T. and Rehmsmeier, M., “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets,” PLOS ONE **10**(3), G. Brock, Ed., e0118432 (2015).

[51] “www.mrmint.fr/outliers-machine-learning.”, .

[52] “www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein/articles/indicateurs-nationaux-de-performance-du-programme-de-depistage-du-cancer-du-sein-sur-la-periode-2015-2016.”, .

[53] Schaffter, T., Buist, D. S. M., Lee, C. I., Nikulin, Y., Ribli, D., Guan, Y., Lotter, W., Jie, Z., Du, H., Wang, S., Feng, J., Feng, M., Kim, H.-E., Albiol, F., Albiol, A., Morrell, S., Wojna, Z., Ahsen, M. E., Asif, U., et al., “Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms,” JAMA Netw. Open **3**(3), e200265 (2020).

[54] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., et al., “International evaluation of an AI system for breast cancer screening,” Nature **577**(7788), 89–94 (2020).

[55] “www.acrdsi.org/Blog/4-Barriers-to-AI-March-2019.”, .

[56] Cannavo, M. J., “www.auntminnie.com/index.aspx?sec=sup&sub=aic&pag=dis&ItemID=127332.”

Annexe

AUTEUR : Nom : LE VOURCH **Prénom : Adrien**

Date de Soutenance : 09/04/2020

Titre de la Thèse : Intelligence artificielle et dépistage du cancer du sein : étude de performance de deux logiciels.

Thèse - Médecine - Lille 2020

Cadre de classement : Radiologie

DES + spécialité : Radiologie et imagerie médicale

Mots-clés : Dépistage – Cancer du sein – Intelligence artificielle

Résumé :

L'avènement des nouvelles technologies basées sur le machine learning va bouleverser la médecine allant du diagnostic, avec un impact certain sur les spécialités d'imagerie médicale, au thérapeutique. De plus en plus de logiciels d'aide au diagnostic, basés sur l'intelligence artificielle sont apparus ces dernières années, notamment dans le domaine de la sénologie et du dépistage du cancer du sein. Le centre hospitalier de Valenciennes, a développé des partenariats avec deux acteurs du marché. Le but de notre étude monocentrique, rétrospective, était d'évaluer la performance de 2 logiciels à trier les examens en normal ou nécessitant de poursuivre les explorations, sur un set issu de données locales de 140 mammographies enrichi en cancer, comparativement à 3 radiologues. Les études étaient considérées comme positives si le BI-RADS initial était supérieur ou égal à 3, les études BI-RADS 1 et 2 étaient considérées comme négatives. Une analyse concomitante a été effectuée en utilisant uniquement les résultats anatomopathologiques des biopsies, où seules les tumeurs prouvées histologiquement étaient considérées comme positives. Les performances des différents opérateurs étaient analysées en fonction de la sensibilité (taux de vrai positif - TVP), du taux de faux positif (TFP), selon les deux catégories de "ground truth". Alors que les performances globales des opérateurs, radiologues et algorithmes, sont comparables sur la première catégorie de "ground truth", l'analyse détaillée par BI-RADS montre des performances médiocres pour un des logiciels sur les BI-RADS extrêmes. Ces mauvais résultats sont confortés par l'analyse des lésions biopsiées. Nous concluons qu'un des deux algorithmes présente des performances inadaptées à une pratique clinique, tandis que le deuxième montre des résultats prometteurs qui devront être confortés par des études de plus grande ampleur sur des data-set de qualité, validés par des radiologues experts.

Composition du Jury :

Président : Monsieur le Professeur Philippe PUECH

Assesseurs : Monsieur le Professeur Olivier ERNST

Monsieur le Professeur Emmanuel CHAZARD

Monsieur le Docteur Nicolas LAURENT

Monsieur le Docteur Luc CEUGNART