UNIVERSITÉ DE LILLE

FACULTÉ DE MÉDECINE HENRI WAREMBOURG

Année : 2021

THÈSE POUR LE DIPLÔME D'ÉTAT

DE DOCTEUR EN MÉDECINE

Impact de l'intelligence artificielle en mammographie de dépistage du cancer du sein

Présentée et soutenue publiquement le 11 octobre 2021 à 16 heures
au Pôle Recherche
par Lan-Anh DANG

_____

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

_____

Avertissement

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

Sigles

ANSM	Agence Nationale de Sécurité du Médicament et des produits de santé
AR	*Augmented radiologist*
AUC	*Area Under Curve*
BDI	Bilan Diagnostic Immédiat
BI-RADS	*Breast Imaging Reporting And Data System*
BRCA	*Breast Cancer (gene)*
CAD	*Computer aided design*
CI	*Confidence interval*
CNAMTS	Caisse Nationale d'Assurance Maladie des Travailleurs Salariés
CRCDC	Centres régionaux de coordination des dépistages des cancers
DGS	Direction Générale de la Santé
DL	*Deep Learning*
DO	Dépistage organisé
FDA	*Food Drug Administration*
HAS	Haute Autorité de Santé
IA	Intelligence artificielle
IC	Intervalle de confiance
IRM	Imagerie par résonance magnétique
ML	*Machine Learning*
NR	*Natural radiologist*
ROC	*Receiver Operating Characteristic*

Sommaire

Résumé

Contexte :

En France, le cancer du sein représente le premier cancer chez la femme et est responsable de plus de 14% de décès par cancer féminin en 2018, avec une prévalence estimée à plus de 7 pour 1000 femmes. Avec la charge de travail croissante des radiologues, comment l'intelligence artificielle (IA) pourrait être une aide dans le dépistage du cancer du sein ?

L'objectif principal de notre étude est de démontrer que les radiologues sont capables de mieux classer les mammographies en catégories BI-RADS, avec le support de l'IA.

Matériel et Méthodes :

Une étude multi-lecteurs, multi-cas a été menée, incluant 314 mammographies. Douze radiologues ont interprété les examens en deux sessions, séparées par 4 semaines de « wash-out », sans et avec l'aide de l'IA. Ils devaient, pour chaque sein de chaque mammographie, marquer la lésion la plus suspecte, et lui attribuer un score BI-RADS « forcé », ainsi qu'un degré de suspicion (allant de 1 à 100, pondéré pour chaque BI-RADS).

Le coefficient de corrélation kappa de Cohen avec pondération quadratique évaluant la concordance inter observateur pour les catégories BI-RADS par sein, l'aire sous la courbe ROC, et les temps de lecture ont été analysés.

Résultats :

En moyenne, le coefficient kappa quadratique a augmenté de manière significative avec le support de l'IA pour l'ensemble des lecteurs (κ = 0.549, 95% CI : [0.528–0.571] sans IA et κ = 0.626, 95% CI : [0.607–0.645] avec IA).

L'AUC s'est significativement améliorée avec l'aide de l'IA (0.739 vs 0.773, p = 0.004).

Le temps de lecture n'a pas été affecté de manière significative pour l'ensemble des lecteurs (106 secondes sans IA et 102 secondes avec IA ; p = 0,754).

Conclusion :

Avec l'aide de l'IA, les radiologues ont mieux classé les mammographies en catégories BI-RADS, sans allonger leur temps d'interprétation.

Introduction

1 Cancer du sein

1.1 Epidémiologie et facteurs de risque

Le cancer du sein représente aujourd'hui le premier cancer chez la femme devant les cancers colorectal et pulmonaire, avec une incidence estimée à environ 58 500 cas, et une mortalité à plus de 12 000 décès par an en 2018 en France métropolitaine.

Il est également responsable du plus grand nombre de décès par cancer chez la femme, avec 14% des décès féminins par cancer en 2018. La prévalence de ce cancer est estimée à environ 7 à 8 pour 1000 femmes. [1,2]

Cette même année, la survie est estimée à 88% à 5 ans dans les zones registres métropolitaines, chez les personnes diagnostiquées entre 2010 et 2015.

Les principaux facteurs de risques connus sont l'âge, la prédisposition génétique, un antécédent personnel de néoplasie mammaire ou d'irradiation thoracique (notamment pour les lymphomes de Hodgkin).

D'autres facteurs de risques ont été identifiés dont certains modifiables liés au mode de vie parmi lesquels on retrouve l'alcool [3] et le tabagisme [4].

1.2 Dépistage du cancer du sein en France

Le dépistage organisé (DO) du cancer du sein a été mis en place en 2004 par la Direction Générale de la santé (DGS), puis généralisé à tout le territoire, suite à une expérience dans 10 départements pilotes menée entre 1989 et 1991 par la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS) [5].

1.2.1 Modalités

Le dépistage organisé du cancer du sein est basé sur une mammographie de dépistage tous les deux ans, et s'adresse aux femmes de 50 à 74 ans, asymptomatiques et sans facteur de risque.

Les femmes reçoivent une invitation à consulter un radiologue agréé dont les coordonnées sont mentionnées dans une pièce jointe au courrier.

Le dépistage comprend une mammographie composée de 2 clichés par sein, (face et oblique externe), ainsi qu'un examen clinique.

Chaque examen se voit attribuer d'un score BI-RADS par sein défini par l'American College of Radiology (ACR) dont la dernière version de 2013 [6] est rapportée en Table 1 :

Table 1. Classification BI-RADS selon l'ACR

Catégorie	Risque de cancer	Conduite à tenir
0 : Investigation incomplète		
1 : Normal	0%	Poursuite de la surveillance classique
2 : Anomalie bénigne	0%	
3 : Anomalie probablement bénigne	> 0% mais < 2 %	Contrôle à 4 ou 6 mois
4. Anomalie nécessitant une biopsie 4A : faiblement suspecte 4B : modérément suspecte 4C : très suspecte	Entre 2 et 95% Entre 2 et 10% Entre 10 et 50% Entre 50 et 95%	Biopsie
5 : Fortement suspecte d'un cancer	> 95%	Biopsie
6 : Cancer prouvé histologiquement	100%	Traitement (Chirurgie…)

Une des spécificités du programme français est la possibilité de réaliser un bilan diagnostic immédiat (BDI) (clichés complémentaires, échographie, biopsie), permettant ainsi de raccourcir le délai de prise en charge de la patiente.

Une deuxième lecture systématique est réalisée par un radiologue expert (L2) lorsque l'examen est normal. Celle-ci permettrait de retrouver 9% des cancers non détectés par le premier lecteur [7].

Les radiologues « premiers lecteurs » ou L1 doivent effectuer au moins 500 mammographies par an, et les deuxièmes lecteurs ou L2 doivent lire au moins 1500 mammographies pour participer au dépistage.

Le contrôle des appareils mammographiques, obligatoire pour tous les appareils (analogiques ou numériques) est placé sous l'autorité de l'agence nationale de sécurité du médicament et des produits de santé (ANSM), et ce deux fois par an [8].

D'un point de vue financier, ce dépistage est pris en charge dans le cadre du tiers payant, donc sans avance de frais, permettant ainsi un accès équitable à l'ensemble des femmes du territoire français.

Le programme français fait appel au structures médicales existantes c'est-à-dire que les mammographies sont réalisées dans les structures d'imagerie médicale publiques et privées.

Ce dépistage était coordonné au niveau départemental par une structure de gestion jusqu'en 2018 ; il l'est désormais au niveau régional par les centres régionaux de coordination des dépistages des cancers (CRCDC), eux-mêmes reposant sur les structures de gestion de leur département.

1.2.2 Population éligible au dépistage organisé

Ce programme national de dépistage organisé s'adresse à plus de 10 millions de femmes de 50 à 74 ans, sans symptôme apparent ni facteur de risque.

Certaines situations imposent un dépistage spécifique du cancer du sein :

- Les femmes à haut risque de cancer du sein : ayant un antécédent personnel de cancer du sein, de carcinome canalaire ou lobulaire in situ, d'hyperplasie canalaire ou lobulaire atypique, patientes ayant subi une irradiation thoracique médicale à haute dose (notamment pour le traitement de la maladie de Hodgkin), patientes présentant certains antécédents familiaux de cancer du sein [9].
- Les femmes à très haut risque de cancer du sein : prédisposition génétique notamment BRCA1 et BRCA 2 [10].

Pour ces patientes, une surveillance spécifique devra être mise en place avec une mammographie annuelle, voire échographie ou IRM.

2 Intelligence artificielle

2.1 Introduction à l'intelligence artificielle

L'intelligence artificielle (IA) existe depuis les années 1950, et est un terme générique désignant les machines effectuant une tâche qui requiert généralement l'intelligence humaine [11].

Le machine learning (ML), ou apprentissage automatique, est apparu quant à lui dans les années 1980, désignant les machines apprenant par expérience et acquérant les compétences sans implication de l'homme ; l'algorithme effectue la tâche de manière itérative, affinant un peu chaque fois pour améliorer le résultat.

Enfin, le deep learning (DL), ou apprentissage profond, sous-catégorie du machine learning, est apparu dans les années 2010, faisant référence au machine learning utilisant des réseaux de neurones, et des algorithmes inspirés du cerveau humain à partir de grandes quantités de data [12]. Le deep learning nécessite un grand nombre de données annotées (requérant du temps médecin, principal facteur limitant). L'annotation consiste à poser un diagnostic et contourer / localiser la lésion.

Le Natural Language Processing (NLP) peut être utilisé pour chercher l'information dans les comptes rendus ; cependant, il faut que l'information soit présente et par conséquent des données peuvent manquer.

Il existe deux types de modèles, les modèles dits « non supervisés », c'est-à-dire qu'à partir de données, on laisse découvrir des classes d'images qui se comportent de la même manière (clustering, modèles descriptifs, données non labellisées), et dont il faudra ensuite comprendre le sens, et les modèles dits « supervisés », c'est-à-dire qu'à partir de données préalablement annotées, on décide à l'avance des classes ou du paramètre continu que l'on souhaite inférer.

L'apprentissage supervisé est le plus utilisé en médecine, se basant sur trois piliers : les données, l'algorithme, et le label = vérité terrain ou « ground truth » en anglais. Premièrement, à partir de données d'apprentissage, et de données annotées,

l'algorithme va mettre au point une équation liant les deux, en prédisant l'annotation à partir des données.

Puis deuxièmement, ce modèle va être appliqué à des nouvelles données, permettant de prédire un résultat à partir de cette équation.

Concernant les données, il faut établir 3 sets :

- Données d'apprentissage pour « entraîner » le modèle, représentatif de la cible avec un équilibre des classes
- Données de validation pour affiner le modèle, permettant une évaluation primaire de l'algorithme entraîné, et permet d'adapter l'algorithme et l'entrainement.
- Et enfin les données test (indépendant) pour tester les performances pour détecter, classer …

Un compromis doit être établi entre précision et généralisabilité ; éviter l'« overfitting » ou surapprentissage, et l'« underfitting » ou sous apprentissage.

Il est possible d'augmenter de manière artificielle le nombre de data par des stratégies de « data augmentation », à coût moindre mais n'équivalant pas à la diversité physiologique.

Par ailleurs, il faut que ces données soient enrichies, qu'elles contiennent suffisamment de cas de chaque classe, et qu'elles représentent la diversité des classes [13].

La vérité des données est établie par le « ground truth » ou vérité terrain.

Pour évaluer la performance du modèle d'intelligence artificielle, l'algorithme calcule une probabilité entre 0 et 1, et en fonction du seuil appliqué, il classifie en présent/absent (détection) ou bénin/malin (classification).

La performance est évaluée par l'aire sous la courbe (AUC ou Area Under Curve) ROC (Receiver Operating Characteristic). Cependant celle-ci n'est pas toujours fiable ; deux courbes ROC avec la même AUC ne sont pas forcément équivalentes ; par ailleurs si l'échantillon est petit, les différences d'AUC peuvent être liées à un seul patient classé différemment.

La courbe ROC permet de choisir le seuil (cut-off) de dichotomisation ou classification. Il faudra privilégier un seuil qui optimise la sensibilité (pour le dépistage par exemple) ou la spécificité (pour le diagnostic, déterminante pour décision thérapeutique).

Les premières applications en imagerie médicale par réseaux de neurones convolutifs ont vu le jour dans le domaine de la dermatologie [14] pour la classification des lésions cutanées (bénignes ou malignes) et de l'ophtalmologie, pour la découverte de facteurs de risque cardiovasculaires à partir de fond d'œil rétinien [15].

De multiples applications de l'IA en radiologie sont possibles : tri des indications, prise de rendez-vous, acquisition des images, reconstruction, radioprotection, CAD (Computer Aided Design), compte rendus standardisés, etc…

Nous nous focaliserons ici sur l'IA en sénologie, et plus particulièrement en mammographie de dépistage.

2.2 Intelligence artificielle et mammographie

Des outils d'aide à la détection (CAD ou Computer Aided Design) ont déjà été utilisés dans les années 2000, notamment aux Etats Unis où ceux-ci étaient validés par la Food and Drug Administration (FDA) et engendrait un coût de plus de 400 millions de dollars par an ; ils ont été finalement remis en question en raison d'une faible spécificité, et d'une diminution de la sensibilité avec l'utilisation des CAD, puis finalement abandonnés en raison de l'absence de bénéfice pour les patientes dépistées. [16,17]

Ces dernières années, l'intelligence artificielle (IA) est en plein essor grâce à l'émergence du Deep Learning (DL) ou apprentissage profond et des réseaux de neurones convolutifs avec le développement de systèmes d'aide à la détection et au diagnostic aux performances bien supérieures à celles des outils précédemment disponibles comme les CAD [18], peu spécifiques. Le deep learning est ainsi une bonne application dans le champ d'action du dépistage du cancer du sein [19].

Pour la détection et la caractérisation lésionnelle en mammographie et tomosynthèse, il existe plusieurs modèles de deep learning à type de CNN (Convolutional Neural Network ou réseaux de neurones convolutifs) à différents niveaux de validation. Ces modèles nécessitent un entrainement sur des milliers de données, sur plusieurs constructeurs afin d'éviter les biais.

Il existe 4 étapes de validation d'un logiciel [20] :

1. Entrainement et validation interne : sur une cohorte enrichie et annotée (étude rétrospective sur data) entrainement et validation externe, sur une cohorte enrichie et annotée, via une étude rétrospective sur data
2. Validation externe (test) : sur une cohorte prévalence supérieure à la normale, comportant des cas subtils (étude rétrospective sur data) : cette étape est cruciale et nécessite la relecture des cas discordants, d'enrichir la base de données en cas positifs, et de développer des outils d'annotation universels
3. Validation externe (test) sur une cohorte avec prévalence égale à la population (étude rétrospective sur data)
4. Validation dans des conditions cliniques (prospectif multicentrique)

De nombreuses applications de l'IA en imagerie mammaire [21] sont actuellement en développement, parmi lesquelles nous pouvons citer :

- Détection du cancer (mammographie, tomosynthèse, échographie, IRM)
- Optimisation des protocoles d'acquisition
- Prédiction du risque

Actuellement 4 logiciels d'IA pour la mammographie sont à l'étude [22]:

- Transpara™ de Screenpoint, à l'étape de validation en prospectif multicentrique (4ème étape)
- Mammoscreen™ de Therapixel à l'étape de validation externe (3e étape, prévalence de la population)

- l'ICad et le Mia à l'étape de validation externe (2e étape, prévalence supérieure à la population).

2.2.1 Modèle d'étude IA versus radiologue

Plusieurs études comparant les performances de la lecture des mammographies de l'IA au radiologue seul ont été réalisées ; une étude compilant l'ensemble des résultats du DREAM Challenge (concours ayant réuni 125 équipes de chercheurs, industriels et start-up, remporté par Therapixel) a permis de démontrer que bien qu'aucun algorithme d'IA seul ne dépassait les performances des radiologues, la combinaison radiologue augmenté par l'IA est meilleure que le radiologue seul mais qu'aucun modèle d'IA seul n'était plus performant que le radiologue [23].

Une autre étude, publiée en 2019, a montré que l'AUC de l'IA était meilleure que la moyenne des 101 radiologues mais moins bonne que les meilleurs radiologues [24].

En 2020, Sasaki et al ont montré que l'AUC de l'IA est inférieure à l'AUC de 3 lecteurs par diminution de la spécificité [25].

Cependant, une publication récente de septembre 2021 dans le BMJ, passant en revue des études utilisant 36 systèmes d'IA, a démontré que ces systèmes étaient moins précis dans 94% des cas qu'un seul radiologue, et que ceux-ci n'étaient pas suffisamment spécifiques pour remplacer la double lecture dans les programmes de dépistage [26].

Le modèle d'étude opposant le radiologue à l'intelligence artificielle semble donc peu pertinent ; il faudrait appréhender cet outil comme une aide au radiologue.

2.2.2 Modèle du radiologue aidé par l'IA

D'autres études, sur le modèle du radiologue « augmenté par l'IA », c'est-à-dire radiologue seul versus radiologue aidé par l'IA ont été menées.

Watanabe et al ont montré en 2019 que certains modèles d'IA étaient une aide au radiologue en augmentant la sensibilité de la détection des cancers du sein [27].

Une étude publiée dans le Lancet Digital Health en 2020 a constaté que l'IA présenterait une meilleure sensibilité dans la détection du cancer pour les masses, les distorsions et les asymétries par rapport aux radiologues ; en outre, elle permettrait une amélioration significative des performances des radiologues avec l'utilisation de leur algorithme, en faveur de son application en mammographie en tant qu'outil d'aide au diagnostic [28].

En terme de densité mammaire, l'IA permet une classification de celle-ci selon le système BI-RADS de l'ACR avec une évaluation précise, standardisée [29] et a une très bonne concordance avec les radiologues séniors et juniors en mammographie digitale et synthétique ainsi que sur le risque de cancer du sein basé sur la densité mammaire [30].

En tomosynthèse, l'IA serait aussi une aide au radiologue en améliorant la détection des cancers avec une réduction du temps de lecture dans l'évaluation des examens [31].

Aujourd'hui, la mammographie synthétique (basée sur la tomosynthèse) tend à être utilisée à la place de la mammographie numérique dans certaines conditions ; une

étude a montré des performances diagnostiques identiques de l'IA entre ces deux clichés [32].

Par ailleurs en terme de santé publique, deux études récentes avec deux constructeurs d'IA différents (Transpara et ICAD) ont montré une diminution des cancers de l'intervalle de 19 et 43 % respectivement [33,34].

L'IA pourrait donc être un bon outil et une aide pour le radiologue dans le dépistage du cancer du sein.

En France, la conduite à tenir dans le cadre du dépistage est basée sur le score BI-RADS : pour un score 1 ou 2, les clichés sont envoyés en 2ème lecture avec poursuite de la surveillance biannuelle ; en dehors de cela, la patiente sort provisoirement du circuit du dépistage organisé : pour un score à 3, il y a indication à une surveillance rapprochée, et pour les scores 4 et 5, une biopsie est indiquée.

L'enjeu dans le dépistage organisé est donc de catégoriser correctement les examens afin d'orienter la prise en charge.

L'objectif principal de notre étude est de démontrer que le radiologue est plus performant pour classer les mammographies en catégories BI-RADS, avec l'aide de l'IA.

Par ailleurs, les objectifs secondaires sont d'évaluer la performance des radiologues sans et avec IA, en termes d'aire sous la courbe ROC, et l'impact de l'IA sur le temps d'interprétation des mammographies.

Article scientifique

Impact of artificial intelligence in breast cancer screening with mammography

Abstract

Objectives:

To demonstrate that radiologists with the help of artificial intelligence (AI), are able to better classify screening mammograms into the correct Breast Imaging Reporting and Data System (BI-RADS) category, and as secondary objectives, to explore the impact of AI on cancer detection and mammogram interpretation time.

Methods:

A multi-reader, multi-case study with cross-over design, was performed, including 314 mammograms. Twelve radiologists interpreted the examinations in two sessions delayed by a 4-weeks wash-out period with and without AI support. For each breast of each mammogram, they had to mark the most suspicious lesion (if any) and assign it with a forced BI-RADS category and a level of suspicion (or “continuous BI-RADS 100”).

Cohen's kappa correlation coefficient evaluating the interobserver agreement for BI-RADS category per breast, and the area under the receiver operating characteristic curve (AUC), were used as metrics and analyzed.

Results: On average, the quadratic kappa coefficient increased significantly when using AI for all readers (κ = 0.549, 95% CI: [0.528–0.571] without AI and κ = 0.626, 95% CI: [0.607–0.6455] with AI).

AUC was significantly improved when using AI (0.74 vs 0.77, p = 0.004).

Reading time was not significantly affected for all readers (106 seconds without AI and vs 102 seconds with AI; p = 0.754).

Conclusion: When using AI, radiologists were able to better assign mammograms with the correct BI-RADS category without slowing down the interpretation time.

Keywords: Artificial Intelligence, Breast cancer, Mammography, BI-RADS classification

Key points:

- AI helps radiologists to assign mammograms with the correct BI-RADS category in breast cancer screening.
- Using the AI support does not slow down the interpretation time.

Abbreviations: AI: Artificial Intelligence; AR: Augmented Radiologist; AUC: Area Under Curve; BIRADS: Breast imaging reporting & data system; CAD: Computer Aided Detection; CI: Confidence Interval; DL: Deep Learning; FDA: Food and Drug Administration; GS: Gold Standard; NR: Natural radiologist ROC: Receiver Operating Characteristic.

Manuscript

1 Introduction

In France, a breast cancer screening program exists for women from 50 to 74 years old without symptoms or family history of breast cancer. If an abnormality is detected, an immediate diagnostic assessment is performed. If the mammogram is assessed as normal, images are sent for second reading to an expert radiologist that can detect up to 9% of cancers.

However, these screening programs are regularly debated due to false positives results leading to unnecessary biopsies, overdiagnosis of non-evolutive cancers leading to unnecessary treatments, psychological impact, and induced x-ray dose [35]. Nevertheless, breast cancer screening program in France has shown a decrease of advanced status cancers [36] leading to a favourable benefit/risk ratio for patients [37].

The use of computer-assisted detection (CAD) tools in mammography for breast cancer screening has been widely studied in the past twenty years, particularly in the United States but abandoned due to the lack of improvement in radiologists' performance [16,17].

In recent years, artificial intelligence (AI) is booming thanks to the appearance of Deep Learning (DL) and Convolutional Neural Networks (CNN) techniques which have led to the development of detection and diagnosis assistance systems with performance superior to those of previously available CAD tools [18].

Numerous applications of AI in breast imaging are currently under development, such as cancer detection for different imaging modalities [21], optimization of acquisition protocols and individual risk prediction.

Concerning mammography, a study has shown a very good agreement on breast density between senior radiologists, junior radiologists and the AI software on 2D and synthetic mammography [30].

Moreover, Rodriguez-Ruiz et al, and Pacilè et al, demonstrated an increase in radiologists' cancer detection performance when using AI without slowing down the reading time [38,39] The use of AI support in screening with tomosynthesis images was found as well to improve cancer detection and on the other hand, reduced reading time [31,40].

Breast cancer is indeed an interesting application as it represents an important public health matter, with more than 58,000 cancers diagnosed in France in 2019 and more than 12,000 deaths per year.

The procedure to be followed is based on the BI-RADS (Breast Imaging-Reporting and Data System) [6] : for 1 and 2 categories, mammograms are sent to second reading system ; if a BI-RADS 3 is assigned, a close monitoring is suggested while for categories 4 and 5 a biopsy is performed.

Therefore, the challenge in organized screening programs is to correctly categorize the examinations to determine the right path to be followed.

Our primary hypothesis is that radiologists are better at classifying mammograms into BI-RADS categories when using AI. The secondary objectives are to evaluate kappa coefficient in 3 categories BI-RADS, radiologist's performances without and with AI in terms of AUC, and the impact of AI on mammography interpretation times.

2 Material and methods

2.1 Data selection and sample size

Data have been retrospective collected from June 2012 to March 2020 at the Valenciennes Hospital (France). Only screening exams were included in the study, acquired with Hologic Selenia® 3D Dimension® system and have been anonymized.

Mammograms and the medical record of the patients were reviewed by an expert radiologist (13 years of experience in breast imaging, 4 years' experience as second reader in French organized screening program), to verify the inclusion criteria.

The expert assigned a forced BI-RADS category (hereafter referred to as "GS" (Gold Standard)) solely based on the 4 standard views mammography and mark the position of the most suspicious lesion (if any). Consequently, for cancer cases visible only on complementary images (magnified views, tomosynthesis), he modified the originally assigned BI-RADS to put himself in the same reading conditions as the radiologists who did the reading sessions (at the risk of under-categorizing the lesion).

Of the 397 examinations in the initial dataset, 329 met the inclusion criteria. Of these, 15 benign cases were randomly excluded to reach the target number of 314 examinations to be included in the study (Table 2). The sample was enriched with cancer cases, but readers were not aware of the proportion. Bifocal cancers were excluded.

Table 2. BI-RADS distribution of the included dataset validated by the expert radiologist (GS)

		Left	Right	Total
Positive cases	BI-RADS 1	1	1	2 (0,3%)
	BI-RADS 2	5	2	7 (1,1%)
	BI-RADS 3	9	2	11 (1,8%)
	BI-RADS 4	29	25	54 (8,6%)
	BI-RADS 5	28	26	54 (8,6%)
	Total	**72**	**56**	**128 (20,4%)**
Negative cases	BI-RADS 1	89	92	181 (28,8%)
	BI-RADS 2	131	142	273 (43,5%)
	BI-RADS 3	18	21	39 (6,2%)
	BI-RADS 4	4	3	7 (1,1%)
	BI-RADS 5	0	0	0 (0%)
	Total	**242**	**258**	**500 (79,6%)**
Total		**314**	**314**	**628**

Two definitions of *ground truth* have been used:

- A standard definition, i.e., cancer cases confirmed by a positive biopsy result and cancer negative verified by a negative follow-up.
- An expert-based definition, i.e., the BI-RADS classification assigned by the expert radiologist during the including phase (GS).

2.2 AI system

The AI software used in this study is Mammoscreen™ v.1.2 created by the French company Therapixel. This software was designed to detect areas suspected of containing breast cancer, to assess their degree of suspicion on 2D digital mammograms.

The system takes as input the cranio-caudal (CC) and mediolateral oblique (MLO) for each breast and provides as output the position of the detected lesions with a suspicion score for each of them ranging from 1 (benign) to 10 (suspicious) by

generating a visual report summarizing the results of the algorithm. The more the score tends towards the extremes (1 or 10), the more the prediction is sure.

The system has been validated for 2D mammography [23] and received Food and Drug Administration (FDA) approval in 2020, as well as CE marking in January 2021.

2.3 Study design

The study was a multi-reader multi-case investigation with cross-over design. There have been two reading sessions delayed by a 4-weeks wash out period. During each session half of the dataset has been read with the AI support (Augmented Radiologist (AR)) and the other half without (Natural Radiologist (NR)). Twelve radiologists were involved in the study: 8 radiologists with more than 3 years of experience (hereafter referred to as “senior”) and 4 radiologists with average experience of 1 year maximum in mammography interpretation (hereafter referred to as “junior”). Reading order was randomized among participants. No additional information such as additional images or information about the patient were available to the readers. Previously acquired mammograms were available for 85 exams out of 314.

Reading time was automatically measured for each case. Readers were aware of the time measurement but blinded to the actual measure.

A training on the AI-tool and its functioning has been carried out before the beginning of the study.

2.4 Reporting

Readers used the usual reading console of the radiology department (5MP screen manufactured by Barco).

For each case, readers were asked to:

- mark the most suspicious lesion per breast on both CC and MLO view (when possible)
- assign a forced BI-RADS score (1 to 5) for each lesion.
- assign a level of suspicion or “continuous BI-RADS 100” defined as follows: a scale ranging from 1 to 100 (1 to 20 for BI-RADS 1, 21 to 40 for BI-RADS 2, and so on)

For exams read with the help of AI, the AI interface was displayed on the reporting console and synchronized with the reading console. In such way, radiologists could check the suspicion score assigned by the AI before reporting their evaluation.

When analysing the results, if readers marked a lesion that was not the correct one (i.e., the mark was beyond 1.5 cm from the centre of the lesion marked by the expert radiologist during the reviewing phase), the case was considered as misclassified.

2.5 Statistical analysis

Sample size has been calculated using the "kappaSize" R package [41] Sample size estimation information has been provided to determine the number of subjects that are required to test the hypothesis H0 : $\kappa = \kappa 0$ vs. H1 : $\kappa = \kappa 1$, at two-sided significance level of 5%, with power of 80%, assuming that the outcome is multinomial with five levels. A minimum of 314 subjects was required to demonstrate an effect on κ of 0.1.

Results are given in terms of Fleiss quadratic kappa correlation coefficient [42] along with their 95% confidence interval (95% CI). The kappa coefficient has been interpreted as follows:

- 0 to 0.4 indicates poor association
- 0.4 to 0.75 indicates medium association
- > 0.75 indicates high association between the two raters.

As there is an inter-observer variability concerning the classification of exams into the BI-RADS 1 and 2 categories as well as into BI-RADS 4 and 5, which both indicate a probably benign and probably malignant examination, respectively, we carried out a second analysis ("3-Cat.BI-RADS") on 3 categories by grouping BI-RADS 1+2, BI-RADS 3 and BI-RADS 4+5.

Secondary endpoints (Quadratic kappa correlation coefficient "3-Cat.BI-RADS", AUC drawn using the "continuous BI-RADS 100" scale and reading time) have been analyzed with statistical methods for reader studies [43].

Statistics tests were bi-sided and considered significant when $p<0.05$.

Data analysis was performed in [44] and NCSS 2021 [45]. For all analyses, the statistical individual is the breast, i.e. 2 breasts per patient, and the most suspicious lesion is described.

3 Results

3.1 Primary endpoint: Impact on agreement between readers (NR and AR) and the expert

3.1.1 Five-categories analysis

Results of the evaluations per breast done by the 12 NR and AR is reported on Table 3.

Table 3. BI-RADS per breast assigned by NR and AR and quadratic kappa

NR (Natural Radiologists)								**Quadratic kappa**
BI-RADS Readers								
BI-RADS Expert (GS)		BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5	TOTAL	0.549, 95% CI : [0.528; 0.571]
	BI-RADS 1*	1255 (57%)	636	198	103	4	2196	
	BI-RADS 2*	801	1819 (54%)	500	238	2	3360	
	BI-RADS 3*	212	167	156 (26%)	62	3	600	
	BI-RADS 4*	196	111	102	260 (35%)	63	732	
	BI-RADS 5*	62	28	37	169	352 (54%)	648	
	TOTAL	2526	2762	993	832	424	7536	
AR (Augmented Radiologists)								
BI-RADS Readers								
BI-RADS Expert (GS)		BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5	TOTAL	0.626, 95% CI : [0.607; 0.645]
	BI-RADS 1*	1357 (61%)	586	198	53	2	2196	
	BI-RADS 2*	819	1901 (57%)	465	172	3	3360	
	BI-RADS 3*	174	178	171 (29%)	68	9	600	
	BI-RADS 4*	178	98	98	276 (38%)	82	732	
	BI-RADS 5*	41	9	35	198	365 (56%)	648	
	TOTAL	2569	2772	967	767	461	7536	

We considered the kappa coefficient with a quadratic weighting, for which a deviation of a single step is given a weight of 1, a deviation of 2 steps is given a weight of 2^2 and so on. This is a severe weighting meaning that it penalizes large deviation very strongly.

The kappa correlation coefficient between the readers and the expert increased significantly from 0.549 [0.528; 0.571] for NR to 0.626 [0.607; 0.645] for AR. (Table 3).

3.2 Secondary endpoints: kappa “3-Cat.BI-RADS” and subgroups based on readers’ experience, ROC, sensitivity, specificity, and reading time

3.2.1 Three-categories analysis “3-Cat.BI-RADS”

We carried out a second analysis grouping categories BI-RADS 1+2, and BI-RADS 4+5.

Table 4. 3-Cat.BI-RADS per breast assigned by NR and AR, quadratic kappa

NR (Natural Radiologists)						**Quadratic kappa**
BI-RADS Readers						
BI-RADS Expert (GS)		BI-RADS 1-2	BI-RADS 3	BI-RADS 4-5	Total	0.528 95% CI : [0.505; 0.550]
	BI-RADS 1-2*	4511 (81%)	698	347	5556	
	BI-RADS 3*	379	156 (26%)	65	600	
	BI-RADS 4-5*	397	139	844 (61%)	1380	
	Total	5287	993	1256	7536	
AR (Augmented Radiologists)						
BI-RADS Readers						
BI-RADS Expert (GS)		BI-RADS 1-2	BI-RADS 3	BI-RADS 4-5	Total	0.614 95% CI : [0.594; 0.635]
	BI-RADS 1-2*	4663 (84%)	663	230	5556	
	BI-RADS 3*	352	171 (29%)	77	600	
	BI-RADS 4-5*	326	133	921 (67%)	1380	
	Total	5341	967	1228	7536	

The quadratic kappa coefficient also increased significantly from 0.528 for NR to 0.614 to AR (Table 4).

3.2.2 Subgroups based on readers’ experience

When evaluating the agreement on subgroups based on readers experience, the kappa coefficient had significantly increased with the help of AI for both subgroups. (Table 5).

Table 5. Quadratic kappa coefficient NR and AR per subgroups of readers based on experience

Juniors		
	NR	**AR**
Quadratic kappa	0.506, 95% CI : [0.466; 0.546]	0.611, 95% CI : [0.575; 0.647]
Seniors		
	NR	**AR**
Quadratic kappa	0.538, 95% CI : [0.511; 0.566]	0.611, 95% CI : [0.585; 0.637]

The performances were calculated considering the histology as ground truth, either biopsy-proven cancer for positive cases, or negative follow-up examination for negative cases. The test was considered positive if the examination was classified by a BI-RADS greater than or equal to 3 (i.e a continuous “BI-RADS100” greater than or equal to 40).

3.2.3 ROC performance (drawn using the continuous “BI-RADS 100 scale”)

On average, radiologists significantly increased their performance in terms of detection with the help of AI, with mean AUC increasing from 0.739 to 0.773 (difference of 0.034; $p = 0.004$) (Table 6, Fig. 1).

The same trend was observed analyzing by experience subgroups (Table 6).

Table 6. Performances (AUC, sensitivity, specificity for a threshold BI-RADS 3), and reading time per reader, NR and AR.

		AUC		
Reader	**NR**	**AR**	**Δ**	
Reader 1	0.747	0.766	0.019	
Reader 2	0.766	0.787	0.021	
Reader 3	0.703	0.758	0.055	
Reader 4	0.725	0.804	0.079	
Reader 5	0.752	0.782	0.030	
Reader 6	0.783	0.757	-0.026	
Reader 7	0.746	0.794	0.048	
Reader 8	0.705	0.727	0.022	
Reader 9	0.731	0.768	0.037	
Reader 10	0.747	0.742	-0.005	
Reader 11	0.720	0.794	0.074	
Reader 12	0.742	0.798	0.056	
Average	**0.739 [0.689. 0.789]**	**0.773 [0.723. 0.823]**	**0.034 [0.012. 0.056]**	***p = 0.004***
Senior	0.744 [0.694, 0.794]	0.776 [0.726, 0.826]	0.032 [0.001, 0.063]	*p = 0.043*
Junior	0.729 [0.671, 0.786]	0.768 [0.710, 0.827]	0.039 [0.010, 0.067]	*p = 0.016*
		Sensitivity		
	NR	**AR**	**Δ**	
Average	0.660 [0.630, 0.700]	0.700 [0.680, 0.720]	0.040 [-0.0002, 0.080]	*p = 0.051*
Senior	0.670 [0.610, 0.720]	0.710 [0.680, 0.740]	0.04 [-0.020, 0.100]	*p = 0.134*
Junior	0.650 [0.580, 0.720]	0.690 [0.650, 0.720]	0.040 [-0.030, 0.100]	*p = 0.227*
		Specificity		
	NR	**AR**	**Δ**	
Average	0.790 [0.740, 0.850]	0.810 [0.770, 0.860]	0.020 [-0.050, 0.090]	*p = 0.570*
Senior	0.790 [0.720, 0.870]	0.810 [0.740, 0.880]	0.020 [-0.080, 0.120]	*p = 0.728*
Junior	0.800 [0.660, 0.930]	0.820 [0.700, 0.940]	0.030 [-0.110, 0.170]	*p = 0.654*
		Reading time		
	NR	**AR**	**Δ**	
Average	106.410 [82.320, 130.520]	101.810 [80.850, 122.760]	-4.620 [-34.730, 25.510]	*p = 0.754*
Seniors	93.970 [66.530, 121.420]	96.170 [67.090, 125.250]	2.200 [-34.080, 38.480]	*p = 0.899*
Juniors	131.300 [69.210, 193.390]	113.080 [65.040, 161.110]	-18.220 [-73.520, 43.070]	*p = 0.490*

.

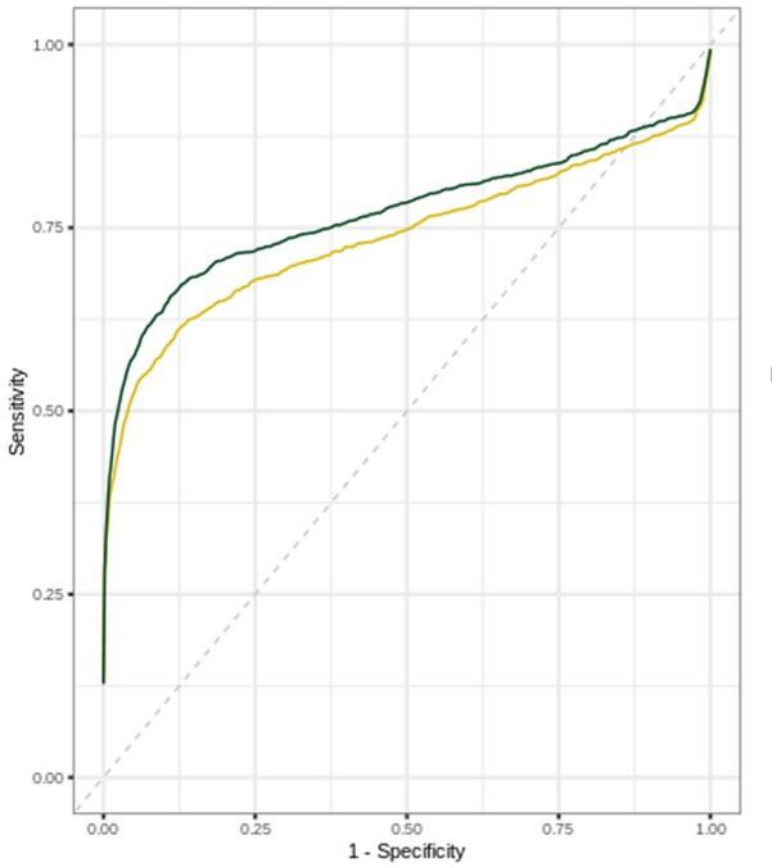


Figure 1. Average ROC curve among readers NR (yellow curve) and AR (green curve).

For two radiologists (one senior and one junior), performance was worst when reading examinations with the AI support with a difference in AUC negative.

3.2.4 Sensitivity and specificity (for BI-RADS greater than or equal to 3)

For example, using a BI-RADS greater than or equal to 3 (i.e a “continuous BIRADS-100” equal to 40), we measured on average, that the sensitivity increased from 0.66 to 0.70, at the limit of significance (p=0.051). Subgroup analysis shows a similar trend of improvement in juniors and seniors (Table 6).

On average, there was no significant difference in specificity without or with AI support. (Table 6).

3.2.5 Reading time

On average, there was no significant difference in mammography interpretation time between reading conditions (NR and AR) (Table 6).

4 Discussion

Considering the 5 BI-RADS categories, radiologists classified better in BI-RADS category for AR. We notice that when grouping in 3 categories (BI-RADS 1+2, 3, 4+5), the quadratic kappa also increased for AR.

The lack of available prior mammograms for most of the examinations included in the dataset, has probably increased the number of potentially benign BI-RADS 3 cases at the expense of the BI-RADS 2 category, and affected the number of false positives.

Actually, applying the BI-RADS mammography classification, high variability has been observed [46], especially for the BI-RADS 3 category [47,48], as we can see in our study (Table 3 and 4).

In addition, it has been observed that the strength of agreement varies widely for different types of mammographic finding, especially for subtle findings such as asymmetries and architectural distortion with a weak agreement [49] which affects BI-RADS categorization.

Moreover, during the validation of the dataset, the expert radiologist had access to all reports and information about the patients but did not have systematically the prior mammogram and validated as BI-RADS 2 some benign lesion described as stable in the patient report. Unfortunately, for technical reasons, prior mammograms could be

made available to readers for a small subset only which may explain certain discrepancies between readers and the expert.

Regarding the secondary endpoints, we were able to demonstrate an improvement in performance for AR by measuring the AUC, which significantly increased.

As an example of these findings, the patient in Figure 2 had a cancer in the left breast. The right breast was cancer-free.

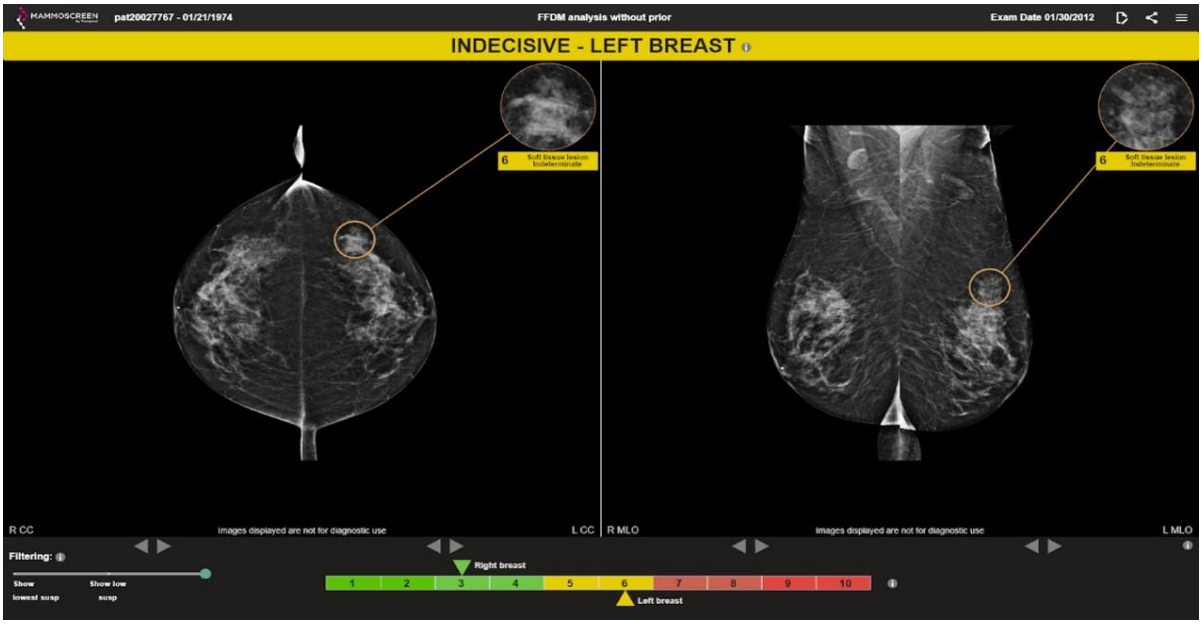


Figure 2. The AI score for this lesion was 6 meaning "indetermined characterization"; without AI, 7 out of 12 readers judged this exam as not suspicious, 2 readers assigned a BI-RADS 3 category, and 3 readers judged the lesion as suspicious for cancer. When reading with AI, one reader only judged the examinations as not suspicious, 2 readers assigned it with a BI-RADS 3 category while 9 readers suspected for cancer

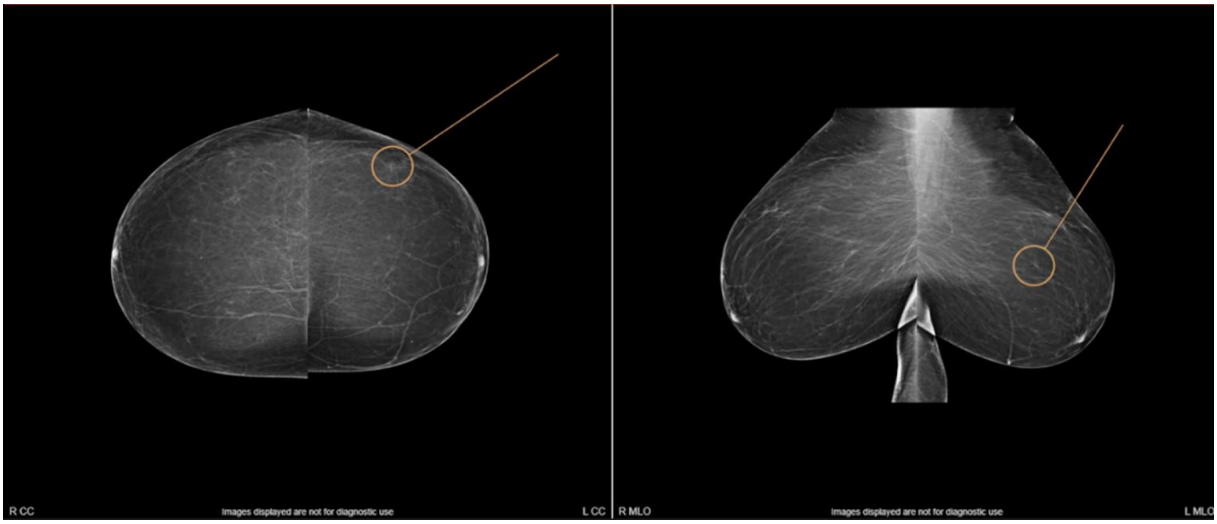
Reading time was weakly affected using IA; we noted however a time gain of 18.13 seconds in the subgroup of junior readers.

The average sensitivity measured in our study (for BI-RADS greater than or equal to 3) seems to be low, that could be explained by the proportion of cancers classified as BI-RADS 1 or 2 validated by the expert and included in the dataset, particularly for cancers only visible on tomosynthesis images (Fig. 3). In fact, only 2D standard 4-view was made available to the readers while in clinical practice additional images

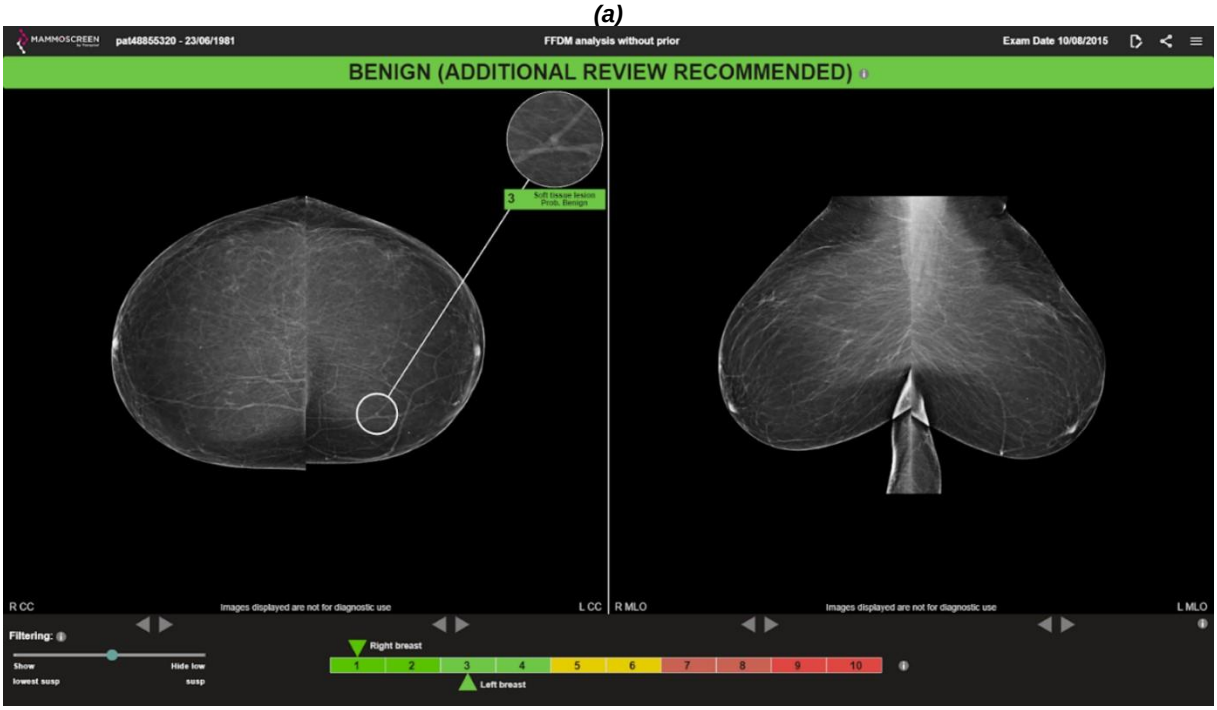
may be considered for the interpretation, notably tomosynthesis images. As an example, for the mammogram in Figure 3, the cancer was visible on tomosynthesis only. The expert radiologist classified this examination as BI-RADS 2, basing his judgement on a mass on the left breast (Fig. 3a). The AI software did not detect any suspicious lesion except for a benign abnormality in the medial region on the left breast (Fig. 3b). All readers classified this examination as benign (i.e., both breasts were assigned with a BI-RADS 1 or 2). A cancer was actually present in the left breast but visible on tomosynthesis only (Fig. 3c). This example shows that AI cannot compensate for the additional imaging performed in clinical practice, particularly tomosynthesis.

Our sample was enriched with more than 20% of cancer cases whereas the natural prevalence is estimated between 7 and 8 per 1000. In addition, there were 11 out of 50 cancerous lesions among BI-RADS 3 (22%), whereas in clinical practice only 2% of them are malignant [6].

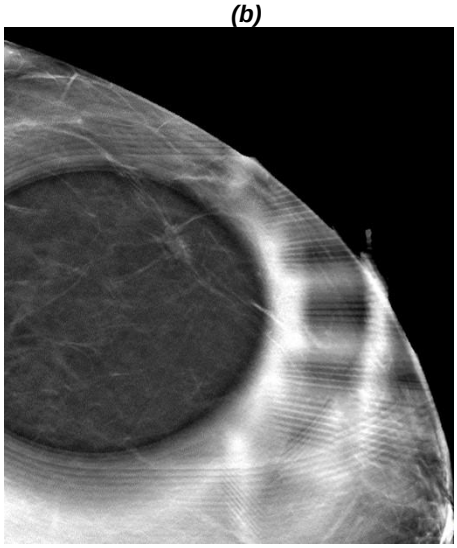
In this type of retrospective study, the reading conditions are far different from clinical practice; the number of examinations read, the reading rate, and the kind of sample are not the same which generate several biases. Gur et al. conducted a study on the laboratory effect by comparing the performance of radiologists in interpreting screening mammograms in clinic with the performance in reading the same examination in laboratory conditions and showed the performance of radiologists was significantly better in clinical conditions [50].



(a)



(b)



(c)

Figure 3. Patient with a cancer on left breast; a) mark by the expert; b) mark by the AI; c) architectural distortion only visible on tomosynthesis

The study in 2020 concerning the same AI tool used in this study, evaluated the performance of radiologists helped by the AI per examination, and showed a better AUC and sensitivity with a decrease in the rate of false negative without affecting readers specificity [39]. Our study demonstrates the interest in using AI in clinical practice for BI-RADS classification of screening mammograms as a help for radiologists, such as it would not be thinkable to replace them, while confirming the improvement in AUC considering a per-lesion analysis.

Our study has several limitations. First, the reference BI-RADS was based on the review of one single expert radiologist. Second, most examinations had no prior mammograms available to the readers. Moreover, readers were using the AI tool for the first time which could have had an impact on the duration of the first interpretations.

This AI-tool have limitations as well: it does not integrate the prior mammograms, tomosynthesis and other clinical information; that could generate false positives and false negatives.

These types of retrospective "laboratory" studies cannot represent performance levels or inter-reader variability during clinical interpretations of the same set of mammograms in a usual work setting. A review of the literature showed that the latest AI models reported good accuracy for breast cancer detection, keeping however methodological biases and weaknesses in the test data limiting its application in a clinical screening setting, needing to be resolved in order to be able to extend AI to large-scale population screening [51]. It would therefore be necessary

to reproduce this same study prospectively in clinical conditions to validate these results.

A European survey on radiologists' opinion on AI, published in February 2021, showed that their perception would influence the adoption of AI in clinical practice and highlights that limited levels of AI-specific knowledge are associated with fear, while intermediate and advanced levels knowledge are associated with a positive attitude towards AI. Additional training could therefore favor the adoption of such tool into clinical practice [52].

In conclusion, radiologists better classified mammograms in BI-RADS categories when reading with the support of AI and improved their AUC without significantly increasing reading times.

Discussion (version française)

En considérant les 5 catégories BI-RADS, les radiologues classaient mieux les mammographies en catégories BI-RADS avec l'aide de l'IA. Nous remarquons qu'en regroupant en 3 catégories BI-RADS (BI-RADS 1+2, 3, 4+5), le coefficient kappa quadratique s'améliorait aussi pour les « radiologues augmentés par l'IA » (AR). Cependant, l'absence de mammographies antérieures disponibles pour la plupart des examens inclus dans la base de données a probablement augmenté le nombre de lésions BI-RADS 3 potentiellement bénignes, aux dépens de la catégorie BI-RADS 2, et affecté le nombre de faux positifs.

De plus, en utilisant la classification BI-RADS, il existe une grande variabilité inter et intra-observateur [46], et particulièrement pour la catégorie BI-RADS 3 [47,48], comme nous pouvons le remarquer dans notre étude (Table 3 et 4).

Par ailleurs, il a été observé que la concordance varie considérablement selon les différents types de lésions mammographiques, en particulier pour les lésions subtiles telles que les asymétries et la distorsion architecturale, pour lesquels la concordance est faible, affectant la classification BI-RADS [49].

Enfin, lors de la validation du dataset, le radiologue expert avait accès à tous les comptes-rendus mais n'avait pas à disposition systématiquement les antériorités, et a donc validé en ACR2 les lésions bénignes décrites dans le compte rendu initial comme stables. Malheureusement, pour des raisons techniques, l'ensemble de ces antériorités n'ont pu être mises à disposition des radiologues lecteurs, pouvant expliquer certaines discordances entre radiologues et expert.

Concernant les critères de jugement secondaires, nous avons pu mettre en évidence une meilleure performance des « radiologues augmentés » (AR) en mesurant une AUC augmentant de manière significative.

Illustrant cette information, la patiente de la mammographie de la Figure 2 présentait un cancer du sein gauche. Le sein droit était indemne de cancer avec un examen de suivi négatif.

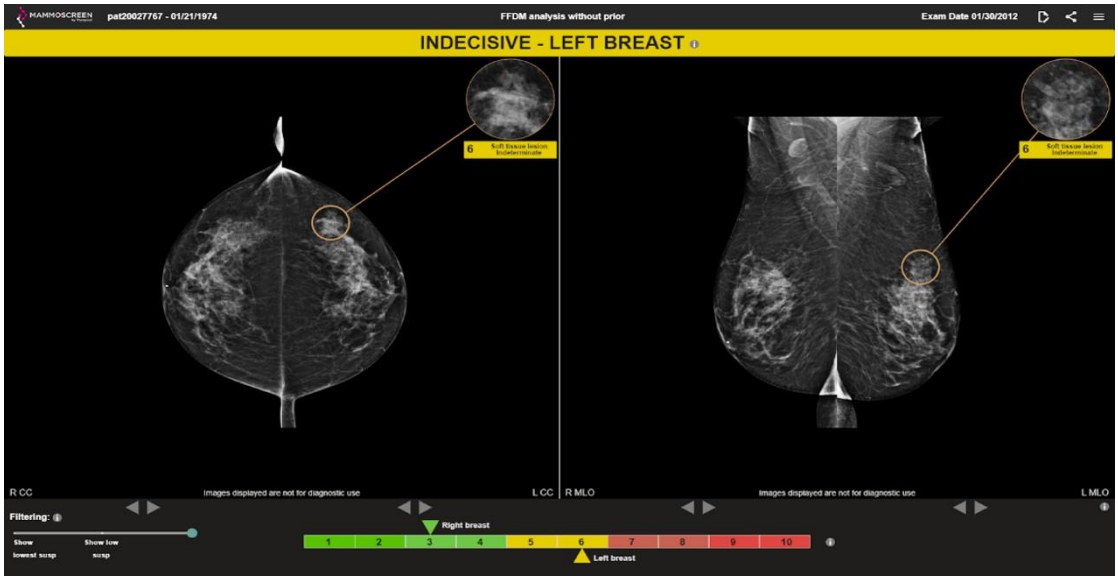


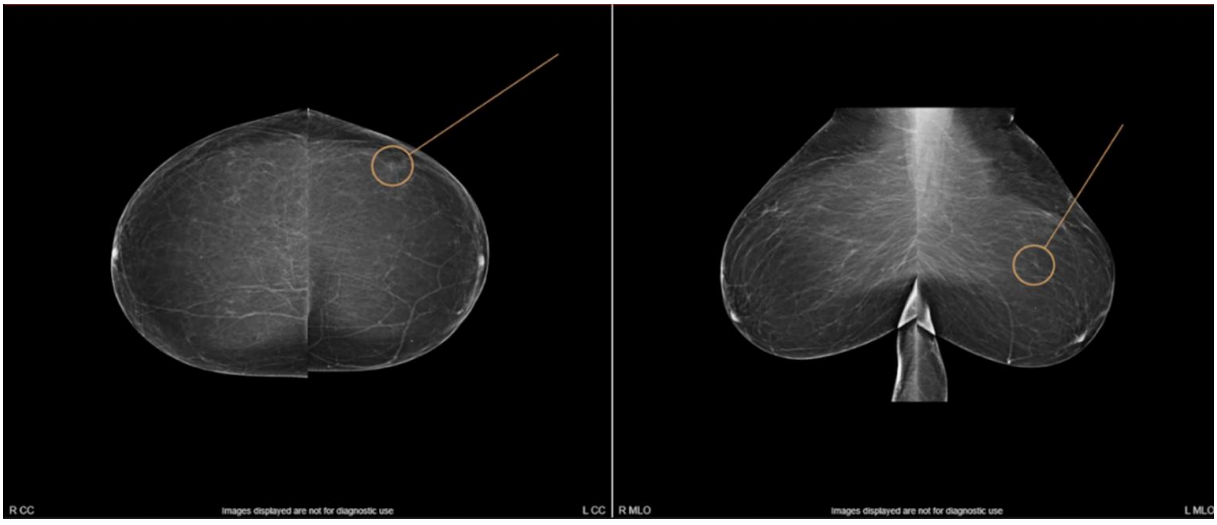
Figure 2. Le score du logiciel d'IA pour cette lésion était de 6 (sur 10), signifiant "caractérisation indéterminée" ; sans IA, 7 radiologues sur les 12 ont considéré cet examen comme non suspect, 2 radiologues ont attribué un score BI-RADS 3, et 3 radiologues suspectaient un cancer du sein gauche. Avec la lecture de l'IA, seulement un radiologue a négativé cet examen, 2 radiologues ont attribué un score BI-RADS à 3, et 9 radiologues suspectaient un cancer au niveau de cette lésion

Les temps d'interprétation étaient faiblement affectés par l'utilisation de l'IA ; on note cependant un gain de temps de 18.13 secondes dans le sous-groupe d'expérience « junior », et une « perte » de temps de 2,2 secondes dans le sous-groupe d'expérience « sénior ».

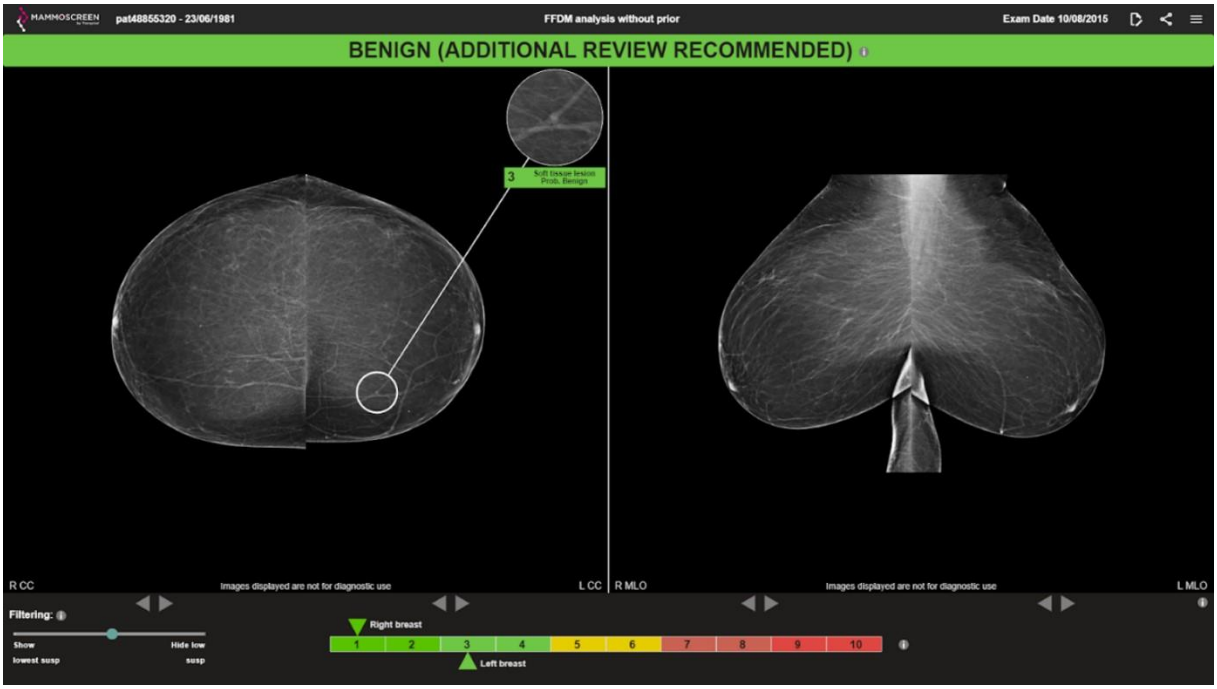
La sensibilité moyenne mesurée dans notre étude (pour un seuil BI-RADS supérieur ou égal à 3 soit un degré de suspicion (ou « BIRADS100 continu » supérieur ou égal

à 40), semble être peu élevée ; ceci pourrait être expliqué par la proportion de cancers classés BI-RADS 1 ou 2, validés par l'expert et inclus dans la base de données, en particulier pour les cancers visibles seulement en tomosynthèse (Fig. 3). En effet, seules les 4 vues standards ont été mises à disposition, alors qu'en pratique clinique, des clichés complémentaires peuvent être réalisés en cas de doute, notamment pour la tomosynthèse.

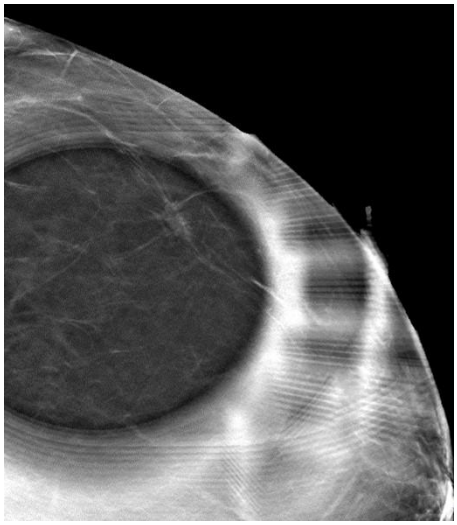
Par exemple, pour la mammographie de la Figure 3, le cancer était visible seulement en tomosynthèse. Le radiologue expert a classé cet examen BI-RADS 2, en basant son jugement sur une masse du sein gauche (Fig. 3a), uniquement sur les 4 vues en 2D. Le logiciel d'IA n'a pas détecté de lésion suspecte, hormis une anomalie bénigne dans la région interne sur le cliché de face gauche (Fig 3b). L'ensemble de ces radiologues a classé cet examen bénin de manière bilatérale. Il existait en fait un cancer du sein gauche, mais visible sous forme d'une distorsion en tomosynthèse seulement (Fig 3c).



(a)



(b)



(c)

Figure 3. Patiente présentant un cancer du sein gauche ; a) marquage de l'expert ; b) marquage de l'IA ; c) Distorsion architecturale visible uniquement en tomosynthèse

Cet exemple montre que l'IA ne peut pallier aux clichés complémentaires réalisés en pratique clinique, notamment la tomosynthèse.

Notre échantillon a été enrichi en cas cancers par rapport aux conditions réelles, avec plus de 20% de cas cancers alors que la prévalence dans la population générale est estimée entre 7 et 8 pour 1000.

En plus des cas cancers classés BI-RADS 1 ou 2, il y avait également dans notre échantillon 11 cas cancers sur 50 lésions BI-RADS 3 soit un taux de 22%, alors qu'en pratique clinique, seules 2% environ sont malignes [6].

Dans ce type d'études rétrospectives, les conditions de lecture sont éloignées de la pratique clinique ; le nombre d'examens lus, le rythme de lecture, l'échantillon ne sont pas les mêmes et engendrent à fortiori plusieurs biais. David Gur et al ont mené une étude sur le « laboratory effect » ou « effet laboratoire » en comparant les performances des radiologues lors de l'interprétation des mammographies de dépistage en clinique avec leurs performances de lecture de ces mêmes examens en condition « laboratoire » d'étude clinique ; celle-ci a démontré qu'en moyenne la performance des radiologues était significativement meilleure en conditions cliniques qu'en conditions « laboratoire » [50].

Une étude utilisant le même outil d'intelligence artificielle Mammoscreen, publiée en 2020 dans Radiology a évalué les performances avec l'aide de l'IA par sein et non par lésion, et mis en évidence une meilleure AUC et sensibilité, avec une diminution du taux de faux négatifs sans affecter leur spécificité [39]. Notre étude démontre l'intérêt de l'utilisation de l'intelligence artificielle en pratique clinique pour la classification BI-RADS des mammographies de dépistage et conforte l'amélioration de l'AUC en considérant l'analyse par lésion (et non par sein).

Cependant notre étude comporte plusieurs limites ; premièrement, le BI-RADS de référence n'a été basé sur la relecture des cas que par un seul radiologue expert. Deuxièmement, la majorité des examens n'avait pas d'antériorité disponible, pour l'expert comme pour les lecteurs.
Par ailleurs les radiologues utilisaient pour la première fois ce logiciel, ce qui a donc pu impacter les durées des premières interprétations.

Les logiciels d'IA ont leurs limites ; ici, le Mammoscreen n'utilise pas les antériorités, la tomosynthèse ou les informations de l'évaluation clinique. Par conséquent certaines lésions détectées peuvent ne pas être pertinentes pour l'utilisateur (faux positifs), et inversement certaines lésions peuvent également être manquées ou sous-évaluées (faux négatifs).

Ce type d'études rétrospectives en « laboratoire » ne peut représenter les niveaux de performance ni la variabilité inter lecteurs au cours des interprétations cliniques du même ensemble de mammographies dans un cadre de travail habituel. Une revue de la littérature menée par Houssami et al en 2019 a montré que les derniers modèles d'intelligence artificielle faisaient état d'une précision généralement bonne pour la détection du cancer du sein, il existait cependant des biais méthodologiques (petites études rétrospectives, dataset contenant une proportion enrichie en cancers ..) et des lacunes dans les données de test (données non représentatives pour un modèle d'entraînement ..) limitant son application dans un contexte clinique de dépistage, devant être résolus afin de pouvoir étendre l'IA au dépistage à grande échelle dans la population [51].

Il faudrait par conséquent reproduire cette même étude de manière prospective dans des conditions cliniques afin de valider ces résultats et s'affranchir de ces biais.

Une enquête européenne sur la vision de l'IA du point de vue du radiologue publiée en février 2021 a montré que la perception des radiologues serait susceptible d'influencer l'adoption de l'intelligence artificielle (IA) dans la pratique clinique. Cette enquête s'intéressait aux connaissances et à l'attitude des radiologues et des internes vis-à-vis de l'IA dans 54 pays pour la plupart européens ; celle-ci met en évidence que des niveaux limités de connaissances spécifiques à l'IA chez les internes en radiologie et les radiologues sont associés à la crainte, tandis que des niveaux de connaissances intermédiaires à avancés spécifiques à l'IA sont associés à une attitude positive envers celle-ci. Une formation supplémentaire pourrait donc améliorer leur intégration dans la pratique clinique [52].

Conclusion

En conclusion, les radiologues ont mieux classé les mammographies en catégories BI-RADS avec l'aide de l'IA et ont augmenté leurs performances, sans temps de lecture additionnel significatif.

Liste des tables

Liste des figures

Références

[1] Cancer du sein 2021. https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein (accessed April 25, 2021).

[2] Le cancer du sein - Les cancers les plus fréquents 2021. https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers/Les-cancers-les-plus-frequents/Cancer-du-sein (accessed April 18, 2021).

[3] Freudenheim JL. Alcohol's Effects on Breast Cancer in Women. Alcohol Res Curr Rev 2020;40:11. https://doi.org/10.35946/arcr.v40.2.11.

[4] Reynolds P. Smoking and breast cancer. J Mammary Gland Biol Neoplasia 2013;18:15–23. https://doi.org/10.1007/s10911-012-9269-x.

[5] Evaluation du programme de dépistage du cancer du sein 2021. https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/articles/evaluation-du-programme-de-depistage-du-cancer-du-sein (accessed September 25, 2021).

[6] American College of Radiology. Breast Imaging Reporting & Data System | American College of Radiology 2013. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads (accessed July 23, 2021).

[7] Dépistage du cancer du sein. Haute Aut Santé 2013. https://www.has-sante.fr/jcms/r_1501534/fr/depistage-du-cancer-du-sein (accessed May 30, 2021).

[8] Le programme de dépistage organisé - Dépistage du cancer du sein 2021. https://www.e-cancer.fr/Professionnels-de-sante/Depistage-et-detection-precoce/Depistage-du-cancer-du-sein/Le-programme-de-depistage-organise (accessed April 25, 2021).

[9] Cancer du sein : modalités spécifiques de dépistage pour les femmes à haut risque. Haute Aut Santé 2019. https://www.has-sante.fr/jcms/pprd_2974673/fr/cancer-du-sein-modalites-specifiques-de-depistage-pour-les-femmes-a-haut-risque (accessed August 12, 2021).

[10] Synthèse - Femmes porteuses d'une mutation de BRCA1 ou BRCA2 / Détection précoce du cancer du sein et des annexes et stratégies de réduction du risque - Ref : RECOBRCASYNTH17 2017. https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Synthese-Femmes-porteuses-d-une-mutation-de-BRCA1-ou-BRCA2-Detection-precoce-du-cancer-du-sein-et-des-annexes-et-strategies-de-reduction-du-risque (accessed August 12, 2021).

[11] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017; 69S:S36–40. https://doi.org/10.1016/j.metabol.2017.01.011.

[12] Posted by Venkatesan M on May 7, 2018, at 9:30pm, Blog V. Artificial Intelligence vs. Machine Learning vs. Deep Learning 2019. https://www.datasciencecentral.com/profiles/blogs/artificial-intelligence-vs-machine-learning-vs-deep-learning (accessed April 25, 2021).

[13] Les données - Développer et valider un algorithme d'IA - Laure Fournier. 2020.

[14] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8. https://doi.org/10.1038/nature21056.

[15] Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018; 2:158–64. https://doi.org/10.1038/s41551-018-0195-0.

[16] Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med 2015;175:1828–37. https://doi.org/10.1001/jamainternmed.2015.5231.

[17] Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007; 356:1399–409. https://doi.org/10.1056/NEJMoa066099.

[18] Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–12. https://doi.org/10.1016/j.media.2016.07.007.

[19] Fernandez-Maloigne C, Guillevin R. L'intelligence artificielle au service de l'imagerie et de la santé des femmes. Imag Femme 2019;29:179–86. https://doi.org/10.1016/j.femme.2019.09.001.

[20] Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 2018 ;286:800–9. https://doi.org/10.1148/radiol.2017171920.

[21] Ceugnart L, Olivier A, Oudoux A. Cancer du sein : la nouvelle imagerie. Presse Médicale 2019;48:1101–11. https://doi.org/10.1016/j.lpm.2019.10.007.

[22] Thomassin-Naggara I, Balleyguier C, Ceugnart L, Heid P, Lenczner G, Maire A, et al. Artificial intelligence and breast screening: French Radiology Community position paper. Diagn Interv Imaging 2019;100:553–66. https://doi.org/10.1016/j.diii.2019.08.005.

[23] Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Netw Open 2020;3:e200265. https://doi.org/10.1001/jamanetworkopen.2020.0265.

[24] Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol 2019;29:4825–32. https://doi.org/10.1007/s00330-019-06186-9.

[25] Sasaki M, Tozaki M, Rodríguez-Ruiz A, Yotsumoto D, Ichiki Y, Terawaki A, et al. Artificial intelligence for breast cancer detection in mammography:

experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer 2020;27:642–51. https://doi.org/10.1007/s12282-020-01061-8.

[26] Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021: n1872. https://doi.org/10.1136/bmj.n1872.

[27] Watanabe AT, Lim V, Vu HX, Chim R, Weise E, Liu J, et al. Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. J Digit Imaging 2019; 32:625–37. https://doi.org/10.1007/s10278-019-00192-5.

[28] Kim H-E, Kim HH, Han B-K, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020;2: e138–48. https://doi.org/10.1016/S2589-7500(20)30003-0.

[29] Ciritsis A, Rossi C, Vittoria De Martini I, Eberhard M, Marcon M, Becker AS, et al. Determination of mammographic breast density using a deep convolutional neural network. Br J Radiol 2019; 92:20180691. https://doi.org/10.1259/bjr.20180691.

[30] Le Boulc'h M, Bekhouche A, Kermarrec E, Milon A, Abdel Wahab C, Zilberman S, et al. Comparison of breast density assessment between human eye and automated software on digital and synthetic mammography: Impact on breast cancer risk. Diagn Interv Imaging 2020; 101:811–9. https://doi.org/10.1016/j.diii.2020.07.004.

[31] van Winkel SL, Rodríguez-Ruiz A, Appelman L, Gubern-Mérida A, Karssemeijer N, Teuwen J, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol 2021. https://doi.org/10.1007/s00330-021-07992-w.

[32] Lee SE, Han K, Kim E-K. Application of artificial intelligence-based computer-assisted diagnosis on synthetic mammograms from breast tomosynthesis : comparison with digital mammograms. Eur Radiol 2021; 31:6929–37. https://doi.org/10.1007/s00330-021-07796-y.

[33] Lång K, Hofvind S, Rodríguez-Ruiz A, Andersson I. Can artificial intelligence reduce the interval cancer rate in mammography screening? Eur Radiol 2021;31:5940–7. https://doi.org/10.1007/s00330-021-07686-3.

[34] Graewingholt A, Rossi PG. Retrospective analysis of the effect on interval cancer rate of adding an artificial intelligence algorithm to the reading process for two-dimensional full-field digital mammography. J Med Screen 2021;28:369–71. https://doi.org/10.1177/0969141320988049.

[35] Gøtzsche PC, Nielsen M. Screening for breast cancer with mammography. Cochrane Database Syst Rev 2006:CD001877. https://doi.org/10.1002/14651858.CD001877.pub2.

[36] Ceugnart L, Rocourt N, Ben Haj-Amor M, Bachelle F, Boulanger T, Chaveron C, et al. [French program of breast cancer screening: Radiologist viewpoint].

Bull Cancer (Paris) 2019; 106:684–92. https://doi.org/10.1016/j.bulcan.2019.03.003.

[37] Coleman C. Early Detection and Screening for Breast Cancer. Semin Oncol Nurs 2017;33:141–55. https://doi.org/10.1016/j.soncn.2017.02.009.

[38] Rodríguez-Ruiz A, Krupinski E, Mordang J-J, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019;290:305–14. https://doi.org/10.1148/radiol.2018181371.

[39] Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool. Radiol Artif Intell 2020;2:e190208. https://doi.org/10.1148/ryai.2020190208.

[40] Conant EF, Toledano AY, Periaswamy S, Fotin SV, Go J, Boatsman JE, et al. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. Radiol Artif Intell 2019;1:e180096. https://doi.org/10.1148/ryai.2019180096.

[41] Rotondi MA. kappaSize: Sample Size Estimation Functions for Studies of Interobserver Agreement. 2018.

[42] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70:213–20. https://doi.org/10.1037/h0026256.

[43] Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. Acad Radiol 2011; 18:129–42. https://doi.org/10.1016/j.acra.2010.09.007.

[44] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

[45] Statistical Software | Data Analysis | Graphics Software | NCSS.com 2021. https://www.ncss.com/software/ncss/ (accessed September 6, 2021).

[46] Boyer B, Canale S, Arfi-Rouche J, Monzani Q, Khaled W, Balleyguier C. Variability and errors when applying the BIRADS mammography classification. Eur J Radiol 2013; 82:388–97. https://doi.org/10.1016/j.ejrad.2012.02.005.

[47] Michaels AY, Chung CSW, Frost EP, Birdwell RL, Giess CS. Interobserver variability in upgraded and non-upgraded BI-RADS 3 lesions. Clin Radiol 2017; 72:694. e1-694.e6. https://doi.org/10.1016/j.crad.2017.03.005.

[48] Ambinder EB, Mullen LA, Falomo E, Myers K, Hung J, Lee B, et al. Variability in Individual Radiologist BI-RADS 3 Usage at a Large Academic Center: What's the Cause and What Should We Do About It? Acad Radiol 2019; 26:915–22. https://doi.org/10.1016/j.acra.2018.09.002.

[49] Lee AY, Wisner DJ, Aminololama-Shakeri S, Arasu VA, Feig SA, Hargreaves J, et al. Inter-reader Variability in the Use of BI-RADS Descriptors for Suspicious Findings on Diagnostic Mammography: A Multi-institution Study of 10 Academic Radiologists. Acad Radiol 2017;24:60–6. https://doi.org/10.1016/j.acra.2016.09.010.

[50] Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The “Laboratory” Effect: Comparing Radiologists’ Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations1. Radiology 2008;249:47–53. https://doi.org/10.1148/radiol.2491072025.

[51] Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI’s potential in breast screening practice. Expert Rev Med Devices 2019; 16:351–62. https://doi.org/10.1080/17434440.2019.1610387.

[52] Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, et al. An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. Eur Radiol 2021 ;31 :7058–66. https://doi.org/10.1007/s00330-021-07781-5.

AUTEUR : Nom : DANG **Prénom : Lan-Anh**

Date de soutenance : Lundi 11 octobre 2021

Titre de la thèse : Impact de l'intelligence artificielle en mammographie de dépistage du cancer du sein

Thèse - Médecine - Lille 2021

Cadre de classement : *Radiodiagnostic et Imagerie Médicale*

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

Mots-clés : Intelligence artificielle, Mammographie, Dépistage, Cancer du sein, Classification BI-RADS

Résumé :

Contexte : En France, le cancer du sein représente le premier cancer chez la femme et est responsable de plus de 14% de décès par cancer féminin en 2018, avec une prévalence estimée à plus de 7 pour 1000 femmes. Avec la charge de travail croissante des radiologues, comment l'intelligence artificielle (IA) pourrait être une aide dans le dépistage du cancer du sein ?
L'objectif principal de notre étude est de démontrer que les radiologues sont capables de mieux classer les mammographies en catégories BI-RADS, avec le support de l'IA.

Matériel et Méthodes : Une étude multi-lecteurs, multi-cas a été menée, incluant 314 mammographies. Douze radiologues ont interprété les examens en deux sessions, séparées par 4 semaines de « wash-out », sans et avec l'aide de l'IA. Ils devaient, pour chaque sein de chaque mammographie, marquer la lésion la plus suspecte, et lui attribuer un score BI-RADS « forcé », ainsi qu'un degré de suspicion (allant de 1 à 100, pondéré pour chaque BI-RADS). Le coefficient de corrélation kappa de Cohen avec pondération quadratique évaluant la concordance interobservateur pour les catégories BI-RADS par sein, l'aire sous la courbe ROC, et les temps de lecture ont été analysés.

Résultats : En moyenne, le coefficient kappa quadratique a augmenté de manière significative avec le support de l'IA pour l'ensemble des lecteurs (κ = 0.549, 95% CI : [0.528–0.571] sans IA et κ = 0.626, 95% CI : [0.607–0.645] avec IA).
L'AUC s'est significativement améliorée avec l'aide de l'IA (0.739 vs 0.773, p = 0.004).
Le temps de lecture n'a pas été affecté de manière significative pour l'ensemble des lecteurs (106 secondes sans IA et 102 secondes avec IA ; p = 0,754).

Conclusion : Avec l'aide de l'IA, les radiologues ont mieux classé les mammographies en catégories BI-RADS, sans allonger leur temps d'interprétation.

<u>Composition du Jury</u> :

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