Brief Report

Comparative Clinical Evaluation of 3 Artificial Intelligence Algorithms for Breast Cancer Screening with Mammography

Alexandra Brion *, Lan-Anh Dang , Edouard Poncelet , Laurie Ferret , Mathilde Vermersch , Laurent Nicolas

Posted Date: 24 January 2024

doi: 10.20944/preprints202401.1706.v1

Keywords: Breast neoplasia; Artificial intelligence; Algorithm; Mammography

Brief Report

Comparative Clinical Evaluation of 3 Artificial Intelligence Algorithms for Breast Cancer Screening with Mammography

Alexandra Brion, Lan-Anh Dang, Edouard Poncelet, Laurie Ferret, Mathilde Vermersch and Nicolas Laurent

Centre Hospitalier de Valenciennes - 114 Av. Desandrouin, 59300 Valenciennes

* Correspondence: alex.brion@free.fr - +33646017902 – 184 rue de Vesle, 51100 REIMS, France

Abstract: Objectives: Few studies have evaluated and compared artificial intelligence (AI) models for breast cancer detection on the same database. The purpose is to compare the performances of 3 artificial intelligence algorithms on the same dataset at a free operative threshold and to evaluate how the performance is impacted by the chosen threshold. **Materials**: In this retrospective single-center study, a dataset of 314 2D screening mammograms performed between 2012 and 2020 was established with a prevalence of 19.6% of histologically proven cancer. Three AI constructors using CE (European conformity) marking algorithms have agreed to take part in the study by submitting the cancer-enriched cohort to their algorithm. They chose a free operative threshold to distinguish benign and malignant mammograms. The most suspicious lesion was marked and correctly located by the AI algorithm. Statistical parameters were analyzed and compared between the algorithms. **Results**: Regarding both sensitivity and specificity, at the chosen threshold, AI 1 had the best compromise between sensitivity (74%) and specificity (79%). AI 2 had a statistically lower sensitivity (52%; $p<0.05$) with a higher specificity (98.4%) than the other AI algorithms. AI 3 had a nonsignificant sensitivity difference (69.9%) but a significantly lower specificity (45.3%; $p<0.001$) than the other two. The performances varied depending on the chosen threshold; when the AI 2 threshold was lowered, the sensitivity increased (69.9%), while a higher specificity (86.1%) was maintained. **Conclusion:** Two AI algorithms stood out in terms of performance when the threshold was optimized, resulting in an acceptable sensitivity and specificity.

Keywords: breast neoplasia; artificial intelligence; algorithm; mammography

Introduction

Breast cancer is the most common cancer diagnosed in women worldwide (1). Prevention and screening have decreased breast cancer mortality by 3 to 35%, depending on the country and the study (2,3). However, mammography screening has limitations such as missed cancers and interval cancers (4), overdiagnosis resulting in overtreatement (5), subjective and variable human interpretation (6,7), and workload challenges (8,9).

Over the last few years, with the advent of deep learning and convolutional neural networks, artificial intelligence (AI) for medical research has advanced. Several studies have evaluated the performance of AI in mammography screening. According to Rodriguez Luiz et al. (10), the performances of AI algorithms were shown to be better than the performance of average radiologists but worse than that of expert radiologists. Additionally, 94% of AI systems were found to be less accurate than radiologists according to a *British Medical Journal* study (11). On the other hand, AI assessment combined with radiologist expertise seems to be more efficient than AI alone or radiologist alone (12,13), as shown by Watanabe et al. (14).

The expected benefits of AI are an overall improvement in radiologist performance, in particular the performance of average readers (10), an aid in the diagnosis of subtle cancers and an improvement in reading time (15). The limits of the use of AI are the absence of consideration of clinical data and previous mammograms (16) and a persistence of false positives leading to overalerting and overdiagnosis (11).

 doi:10.20944/preprints202401.1706.v1

To our knowledge, only one independent Swedish study has evaluated the performance of 3 commercialized artificial intelligence algorithms as independent mammography readers compared to radiologists. That study showed that one of the AI algorithms was more efficient than radiologists alone (17).

Currently, in Europe, several companies offer breast cancer detection software that has obtained a CE (European conformity) marking. In clinical use, AI systems target lesions and suggest cancer risk categories, generally based on three levels of risk. The intermediate category often concentrates a large proportion of false-positive lesions that may affect radiologist interpretation. It would seem necessary to improve this categorization by determining a threshold that would allow effective detection while limiting false positives. Furthermore, locating the lesions is rarely required.

The main aim of the current study was to compare the performance of 3 AI algorithms based on the same dataset at a free operative threshold. The second outcome was to evaluate performance variations when modifying the operating threshold.

Material and methods

Data selection and sample size

The study sample was extracted retrospectively from a dataset of Valenciennes Hospital (France). The oldest exam dates back to June 2012 and the most recent from March 2020. All mammograms were acquired with a Hologic Selenia 3D Dimension® system. The dataset only included screening exams. All patients were eligible for this retrospective institutional review board (IRB)-approved study (IRB number, CRM-2304-335). Written informed consent was waived by IRB.

The included women were aged from 40 to 74 years, asymptomatic, without any history of personal breast cancer and had a complete screening examination prior to diagnosis. Women with multifocal cancers were excluded.

Mammograms and patients' medical records were reviewed by an expert radiologist with 15 years of experience in breast imaging and 6 years of experience as a second reader in the French organized screening program. He checked the inclusion criteria to mark suspicious cancer lesions and assigned a BI-RADS score (American College of Radiology classification) (18).

A total of 314 bilateral mammograms that met the inclusion criteria were randomly selected to be included in the study. The sample was enriched in cancer cases, reaching a prevalence of 19.6% (Table 1).

Table 1. Dataset Distribution.

Variables	N (%)
Status	
Benign	505 (80.4)
Malignant	123 (19.6)
Location of malignant lesion	
Left	53 (43.1)
Right	36 (29.3)
Both	34 (27.6)
Cancer type	
Mass	74 (60.2)
Calcification	32 (26)
Focal asymmetry	7 (5.7)
Architectural distorsion	10 (8.1)
Cancer BI-RADS score	
2	2 (1.6)
3	11 (8.9)
4	55 (44.7)
5	55 (44.7)

 doi:10.20944/preprints202401.1706.v1

Breast density	
A	57 (18.2)
B	164 (52.2)
C	90 (28.7)
D	3 (0.96)

Malignant cases were composed of 60.2% masses, 26% calcifications, 5.7% focal asymmetries and 8.1% architectural distortions. 89.4% of cancerous lesions were initially classified as BI-RADS 4 (suspicious) or 5 (highly suggestive of malignancy).

The gold standard was defined by histology, i.e., a positive biopsy for cancer cases and a 2-year negative control mammogram for noncancer cases.

AI system

Among six available AI programs with 2D models, three agreed to take part in the study (Incepto, Therapixel, Hera-Mi) and three did not wish to participate or were unable to do so for technical or logistical reasons (I-CAD, Hologic, Lunit).

The following AI programs were used in the study :

-Transpara v.1.7.3 from the French company Incepto© and developed by the Dutch company Screenpoint©; this program circumscribes the lesion and gives a region score from 1 to 98 and a global risk category per patient: low risk if the region score is between 1 and 43, intermediate risk if the score is between 43 and 75 and high risk if the score is between 75 and 98.

-Mammoscreen™ v.1.2 from the French company Therapixel©; this program targets the lesion and gives a malignancy score on a scale from 1 to 10 per lesion and per breast. Three categories are identified: low risk from 1 to 4, intermediate risk from 5 to 7 and high risk from 8 to 10.

-Breast-SlimView® v1.8.0 from the French company Hera-Mi©; this program generates a synthetic image by blurring the normal breast to highlight the suspect zone.

All of these programs have received a CE marking.

For statistical analysis, the AIs were anonymized and randomly named AI 1 to 3.

Study design

The 314 anonymized mammograms from the dataset were used by AI constructors in January 2023. Each mammogram was the result of 2-view full-field digital mammography (FFDM) of each breast without tomosynthesis or any other clinical information.

The dataset was available on a shared space with secure access for AI algorithm processing. The results were returned within 48 hours of the download link being sent.

Each algorithm defined an optimal cancer detection threshold for screening to distinguish the positive and negative lesions. A lesion was considered positive if its score was above the threshold. If several lesions were positive on the same breast, only the most suspect lesion was considered. A summary spreadsheet, reported the results of each mammographic incidence, including the label (1 if there was a positive lesion, 0 if not), a probability score between 0 and 1 and coordinates of the positive lesion. A screenshot of each case was also provided with the positive lesion marked.

Data analysis

A reviewer processed the data verification and analysis.

Constructor spreadsheets were compared to the ground truth file with the aid of an informatic script in Python code. The coordinates of the positive lesions were also validated with the script with a 15 mm deviation tolerance from the center of the marked lesion.

The results were verified with the screenshots provided.

The results were reported in tables by transcribing the number of true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs).

 doi:10.20944/preprints202401.1706.v1

Analysis was evaluated per breast, i.e., cancer lesions were considered to be correctly classified if they were correctly marked on at least one occurrence for each breast.

Statistical analysis

Categorical data are represented as numbers (percentage). Continuous variables are represented as the median (range). All statistical analyses were blinded. The diagnostic performance of each dataset was evaluated on a per breast basis using sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and accuracy (Acc). These features were compared using a two-sided chi-square test. All statistical analyses were performed using SPSS version 23 and MedCalc. A p value < 0.05 was considered significant. Balanced accuracy was calculated as the average of sensitivity and specificity in the case of unbalanced classes.

Results

Primary endpoint: Comparison of the performances of the 3 AI algorithms

Table 2 reports the diagnostic metrics of cancer detection for each AI algorithm, including sensitivity, specificity, PPV, NPV, accuracy, and balanced accuracy.

Table 2. Statistical Metrics for each AI.

Breast analysis	AI 1 (%)	AI 2 (%)	AI 3 (%)
SE	91/123 (74)	64/123 (52)	86/123 (69.9)
SP	399/505 (79)	497/505 (98.4)	229/505 (45.4)
PPV	64/72 (46.2)	64/72 (88.9)	86/362 (23.8)
NPV	399/431 (92.6)	497/556 (89.4)	229/266 (86.1)
Accuracy	490/628 (78)	561/628 (89.3)	315/628 (50.2)
Balanced accuracy	76.5	75.2	57.6

The performances of the algorithms were heterogeneous; AI 1 had an acceptable sensitivity (74%), specificity (79%) and accuracy (78%), whereas AI 2 had a high specificity (98.4%) and accuracy (89.3%) but lower sensitivity (52%). Overall, AI 3 had lower results except for sensitivity (69.9%) (Table 2).

On a per breast analysis (Table 3), the performances of the 3 AIs were compared two by two.

Table 3. AI Algorithms Compared Two by Two with p Values.

Breast analysis (%)	AI 1	AI 2	AI 3	p value		
				AI 1 vs AI 2	AI 1 vs AI 3	AI 2 vs AI 3
SE	74	52	69.9	< 0.001	0.478	0.004
SP	79	98.4	45.4	< 0.001	< 0.001	< 0.001
PPV	46.2	88.9	23.8	< 0.001	< 0.001	< 0.001
NPV	92.6	89.4	86.1	0.086	0.005	0.168
Accuracy	78	89.3	50.2	0.001	0.004	< 0.001

At the chosen threshold, differences in sensitivity between AI 2 (52%) and the other two algorithms (AI 1 (74%) and AI 3 (69.9%)) were statistically significant (p<0.001 and p=0.004, respectively), whereas no significant difference between AI 1 and AI 3 was demonstrated ($p = 0.478$).

The specificity of AI 2 (98.4%) was significantly superior to that of AI 1 (79%) and AI 3 (45.4%), as was the difference between AI 1 and AI 3.

The positive predictive value of AI 2 (88.9%) was also significantly better than that of AI 1 (46.2%) and AI 3 (23.8%).

There was no significant difference in the negative predictive value between AI 1 and AI 2.

 doi:10.20944/preprints202401.1706.v1

Accuracy was significantly higher with AI 2 (89.3%) compared to AI 1 (78%), which itself was higher than AI 3 (50.2%). When classes were balanced, the accuracy of AI 2 decreased from 89.3% to 75.2%, approximately equivalent to AI 1 (76.5%).

Secondary endpoints: Analysis of performances by varying the operating threshold

Based on the scores for AI 2 for each lesion, we secondarily analyzed performances on a per breast analysis by varying the threshold (Table 4). The lower the threshold was, the greater the sensitivity increased (69.9% with a 0.5 threshold and 74.8% with a 0.4 threshold), reaching sensitivity values statistically comparable to AI 1 and AI 3 ($p>0.05$). On the other hand, specificity decreased but remained significantly superior to that of AI 1 at the threshold of 0.5 (86.1%; $p<0.05$). Balanced accuracy remained at a high level, reaching 79.2% at the 0.6 threshold and 78% at the 0.5 threshold.

Table 4. AI 2 Performance based on Various Thresholds.

Threshold	0.8	0.7	0.6	0.5	0.4
SE (%)	52	59.4	66.7	69.9	74.8
SP (%)	98.42	96.2	91.7	86.1	70.5
PPV (%)	89	79.4	66.1	55.1	38.2
NPV (%)	89	90.7	91.9	92.2	92
Accuracy (%)	89.3	89	86.8	83	71.3
Balanced accuracy (%)	75.2	77.8	79.2	78	72.7

Discussion

The main outcome of this study revealed significant differences between AI systems depending on the intrinsic performance of algorithms and on the threshold chosen.

If we consider all the performance parameters of breast cancer screening algorithms at the initially chosen threshold, AI 1 achieved a compromise between an acceptable detection rate (Se=74%) and a correct specificity (Sp=79%), generating a moderate PPV (46.2%) that would partly limit overalerting.

In contrast, AI 2 had excellent accuracy, based on its high specificity (98.4%) and PPV (88.9%), but a low sensitivity in the context of cancer screening (52%).

Finally, AI 3 had a moderate sensitivity (69.9%), but its low specificity (45.4%) and PPV (23.8%) had a definite impact on the number of false positives and thus led to overalerting. Indeed, these false-positive lesions restricted radiologist interpretation and could lead to overmedicalization, loss of time and decreased radiologist confidence in the system in clinical practice.

Overall, the sensitivity rates of the AI algorithms appeared to be lower compared to the literature data, ranging from 67% to 81.9% at a fixed specificity rate of 96.6% in the Swedish study (18) and 96.2% compared to a radiologist specificity of 66.9% in Lotter's study (19), However, a retrospective study published in *Nature* found lower sensitivities for AI systems (56%) for a specificity of 84% (20), which is comparable to our study.

These rates could be explained by the fact that the dataset came from an expert center with several cancers that were difficult to detect, such as two cancers visible only on tomosynthesis, initially classified as BI-RADS 2, and 11 cancers classified as BI-RADS 3. Moreover, the correct location of lesions compared to those marked by the expert radiologist was required whereas in recent studies, the analysis was only performed per positive mammogram.

Sensitivity depends on false negatives. In our study, they can be distinguished into two categories. The first category consists of lesions that were detected by the AI algorithms and classified as negative because their score was below the chosen threshold, as shown in Figure 1, where 2 AI algorithms marked the right lesion but were considered misclassified. The second category includes lesions that were not detected by the algorithm, as shown in Figure 2, where none of the algorithms found the right lesion.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 January 2024 doi:10.20944/preprints202401.1706.v1

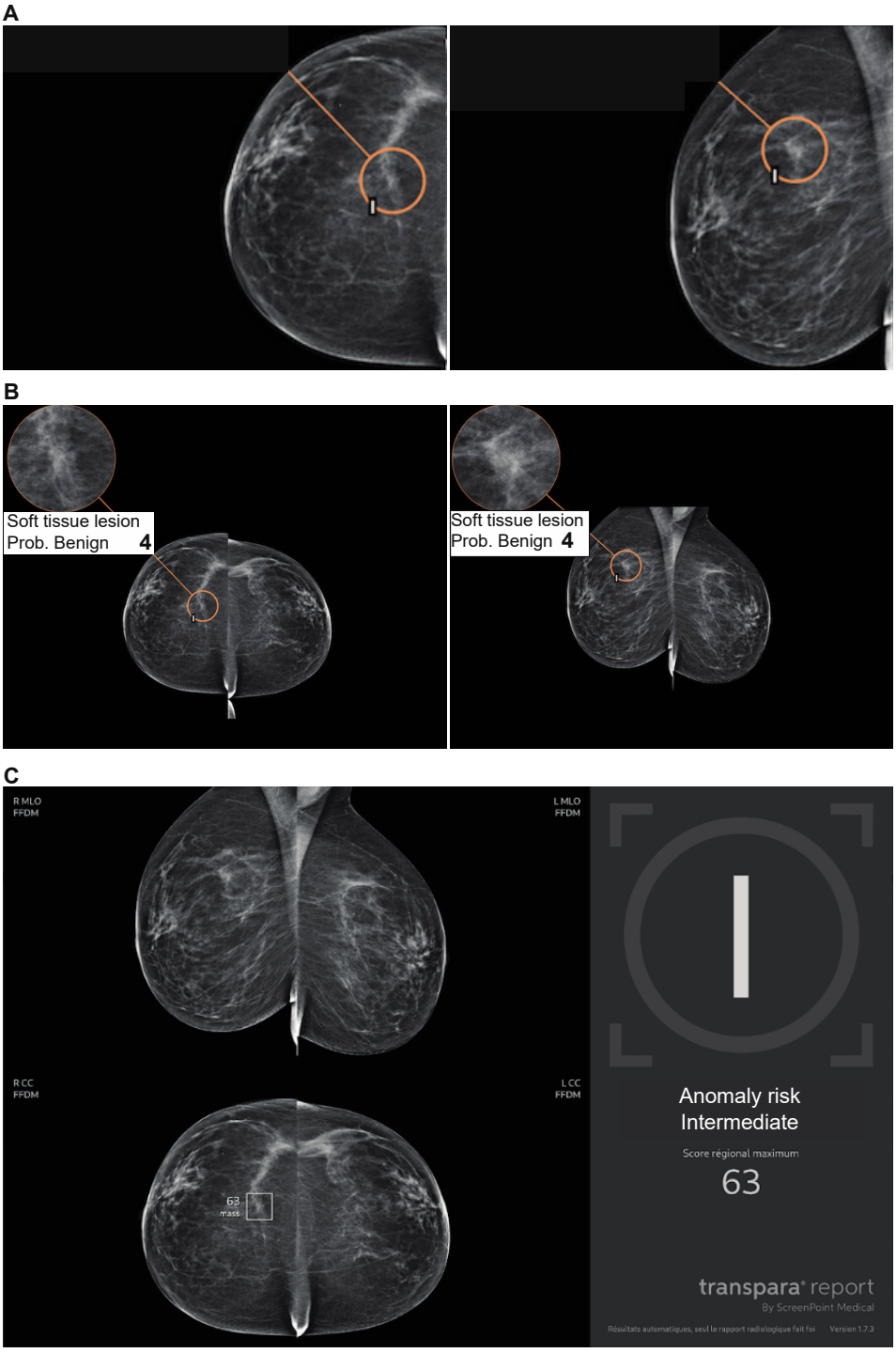

Figure 1. Distortion marked as BI-RADS 4 by the expert radiologist. (a) False negative for 2 AI algorithms that marked the proven lesion, but the score was under the threshold (b, c), and no mark for the third AI algorithm.

 doi:10.20944/preprints202401.1706.v1

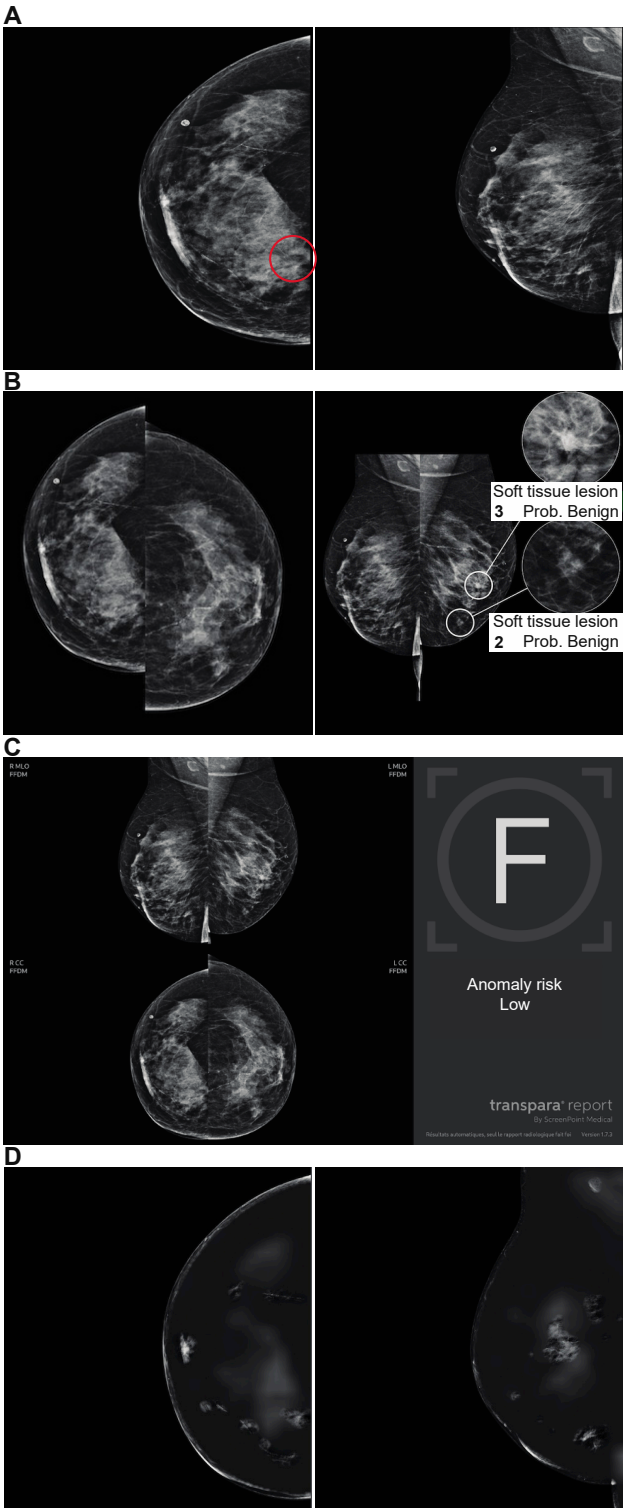

Figure 2. Subtle cancer only visible on one incidence marked as BI-RADS 4 by the expert radiologist. (a) False negative for the 3 AI algorithms, which did not mark the proven lesion (b, c, d).

While most studies impose a fixed value of sensitivity or specificity to compare algorithm performances (10,19,21), the choice of the "optimal" threshold for each algorithm was left to the constructors. To choose it, constructors aim for a recall rate that is approximately 20 to 40% for a cancer-enriched cohort. Currently, the recall rate in the United States is more than approximately 10% in a screening population (22).

The choice of the threshold affects AIs performances. By lowering AI 2's threshold, the sensitivity was significantly improved, reaching values without a significant difference with AI 1 at

 doi:10.20944/preprints202401.1706.v1

the threshold of 0.5 (Table 4). After a discussion with AI 2's constructor, retrospectively, the correct threshold should have been lower than the chosen one, to have the best compromise between sensitivity and specificity. This example illustrates the impact of the threshold choice. We understand that a "perfect model" that would detect all cancers without generating any false-positive results does not exist.

The strength of our study is that we conducted a clinical study of several AI solutions currently marketed, according to similar exercise modalities (with respect to the 48-hours deadline for the return of results) with verification of the correctness of the results by the screenshots provided.

However, this was a retrospective and single-center study, with a dataset enriched in cancers and an analysis only of 2D mammograms. The limitations are the low number of cases in the database and an artificial high prevalence due to an a cancer-enriched cohort which differs from an actual screening setting. Furthermore, only 2D AI models have been evaluated while studies have shown the usefulness of tomosynthesis, which increase the number of cancers detected in the radiologist's clinical practice (23) and the importance of medical history (17). AI algorithms are currently being developed with the integration of 3D analysis and prior mammograms, which will require further studies.

At present, the use of AI tools for breast cancer screening is intended more for centers performing routine mammography as an aid to vigilance. They could improve practices and performance depending on the expertise and experience of the radiologist, which may vary from center to center. However, the added value for an expert center remains to be demonstrated.

Real-life conditions may lead to new biases such as the automation bias of radiologists when using an AI system that could influence decisions made by radiologists, leading to an impairment of performance (24). This tendency had already been noticed with the use of CADs, which decreased the sensitivity of radiologists (25,26).

Despite promising performance, the place of AI models in patient care remains to be determined, as highlighted in the state of the art by Sechopoulos et al. (27) based on large-scale prospective studies at several centers under actual screening conditions (28). Recent prospective studies seem to demonstrate the non-inferiority of AI-supported screening compared with a standard double reading, with lower workload (29). This raises the question of the evolution of screening towards the replacement of the second reading.

Conclusion

In conclusion, the current study revealed heterogeneous results for 3 AI algorithms depending on their intrinsic performance and their chosen operating threshold. Two AI algorithms stand out in terms of performance when the choice of threshold is optimized to obtain the best compromise between the greatest number of cancers detected and the least number of false positives to be treated.

This study has thus highlighted the importance of the choice of the threshold and its statistical implications. Radiologists should be aware of this issue before implementing an AI solution in clinical practice so that the chosen AI system can provide assistance to users.

List of abbreviations

AI – Artificial intelligence
BI-RADS - Breast Imaging-Reporting And Data System
CAD - computer-aided detection system
CE – European conformity
FFDM - Full-field digital mammography
IRB - Institutional review board
NPV – Negative predictive value
PPV – Positive predictive value
Se – Sensitivity
Sp - Specificity

References

1. Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast. déc 2022;66:15-23.
2. Broeders M, Moss S, Nyström L, Njor S, Jonsson H, Paap E, et al. The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies. J Med Screen. 2012;19 (Suppl 1):14-25.
3. SPF (santé publique France). Éthique et dépistage organisé du cancer du sein [Internet]. [cité 18 juill 2023]. Disponible sur: https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Ethique-et-depistage-organise-du-cancer-du-sein-Synthese
4. SPF. Sensibilité et spécificité du programme de dépistage organisé du cancer du sein à partir des données de cinq départements français, 2002-2006. Numéro thématique. Dépistage organisé du cancer du sein. https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein/sensibilite-et-specificite-du-programme-de-depistage-organise-du-cancer-du-sein-a-partir-des-donnees-de-cinq-departements-francais-2002-2006.-nume. Accessed February 27, 2023.
5. Paci E. Summary of the evidence of breast cancer service screening outcomes in Europe and first estimate of the benefit and harm balance sheet. J Med Screen 2012;19 Suppl 1:5-13. doi: 10.1258/jms.2012.012077
6. Elmore JG, Jackson SL, Abraham L and al. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 2009;253:641-651. doi: 10.1148/radiol.2533082308
7. Miglioretti DL, Smith-Bindman R, Abraham L and al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. J Natl Cancer Inst 2007;99:1854-1863. doi: 10.1093/jnci/djm238
8. Giess CS, Wang A, Ip IK, Lacson R, Pourjabbar S, Khorasani R. Patient, radiologist, and examination characteristics affecting screening mammography recall rates in a large academic practice. J Am Coll Radiol 2019;16:411-418. doi: 10.1016/j.jacr.2018.06.016
9. Barlow WE, Chi C, Carney PA and al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96:1840-1850. doi: 10.1093/jnci/djh333
10. Rodriguez-Ruiz A, Lång K, Gubern-Merida A and al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019;111:916-922. doi: 10.1093/jnci/djy222
11. Freeman K, Geppert J, Stinton C and al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021;374:n1872. doi: 10.1136/bmj.n1872
12. Dang LA, Chazard E, Poncelet E and al. Impact of artificial intelligence in breast cancer screening with mammography. Breast Cancer 2022;29:967-977. doi: 10.1007/s12282-022-01375-9
13. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, Mann RM. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290:305-314. doi: 10.1148/radiol.2018181371
14. Watanabe AT, Lim V, Vu HX and al. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. J Digit Imaging 2019;32:625-637. doi: 10.1007/s10278-019-00192-5
15. Balleyguier C, Arfi-Rouche J, Levy L, Toubiana PR, Cohen-Scali F, Toledano AY, Boyer B. Improving digital breast tomosynthesis reading time: a pilot multi-reader, multi-case study using concurrent computer-aided detection (CAD). Eur J Radiol 2017;97:83-89. doi: 10.1016/j.ejrad.2017.10.014
16. Choi WJ, An JK, Woo JJ, Kwak HY. Comparison of diagnostic performance in mammography assessment: radiologist with reference to clinical information versus standalone artificial intelligence detection. Diagnostics (Basel) 2022;13:117. doi: 10.3390/diagnostics13010117
17. Salim M, Wåhlin E, Dembrower K and al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020;6:1581-1588. doi: 10.1001/jamaoncol.2020.3321
18. ACR (American college of Radiology). Breast Imaging Reporting & Data System [Internet]. [cité 29 mars 2023]. Disponible sur: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads
19. Lotter W, Diab AR, Haslam B and al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med 2021;27:244-249. doi: 10.1038/s41591-020-01174-9

20. McKinney SM, Sieniek M, Godbole V and al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89-94. doi: 10.1038/s41586-019-1799-6
21. Schaffter T, Buist DSM, Lee CI and al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3:e200265. doi: 10.1001/jamanetworkopen.2020.0265
22. Rauscher GH, Murphy AM, Qiu Q, Dolecek TA, Tossas K, Liu Y, Alsheik NH. The "sweet spot" revisited: optimal recall rates for cancer detection with 2D and 3D digital screening mammography in the metro Chicago breast cancer registry. AJR Am J Roentgenol 2021;216:894-902. doi: 10.2214/ajr.19.22429
23. McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. JAMA Oncol 2016;2:737-743. doi: 10.1001/jamaoncol.2015.5536
24. Dratsch T, Chen X, Mehrizi MR and al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023;307:e222176. doi: 10.1148/radiol.222176
25. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175:1828-1837. doi: 10.1001/jamainternmed.2015.5231
26. Chan HP, Samala RK, Hadjiiski LM. CAD and AI for breast cancer-recent development and challenges. Br J Radiol 2020;93:20190580. doi: 10.1259/bjr.20190580
27. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: state of the art. Semin Cancer Biol 2021;72:214-225. doi: 10.1016/j.semcancer.2020.06.002
28. Dembrower K, Crippa A, Colón E, Eklund M, Strand F, ScreenTrustCAD TrialConsortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. Lancet Digit Health. sept 2023;S2589-7500(23)00153-X.
29. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. août 2023;24(8):936-44.