



OPEN

## Saliency of breast lesions in breast cancer detection using artificial intelligence

Said Pertuz<sup>1</sup>, David Ortega<sup>1</sup>, Érika Suarez<sup>1</sup>, William Cancino<sup>1</sup>, Gerson Africano<sup>1</sup>, Irina Rinta-Kiikka<sup>2,3</sup>, Otso Arponen<sup>2,3</sup>, Sara Paris<sup>4</sup> & Alfonso Lozano<sup>4</sup>

The analysis of mammograms using artificial intelligence (AI) has shown great potential for assisting breast cancer screening. We use saliency maps to study the role of breast lesions in the decision-making process of AI systems for breast cancer detection in screening mammograms. We retrospectively collected mammograms from 191 women with screen-detected breast cancer and 191 healthy controls matched by age and mammographic system. Two radiologists manually segmented the breast lesions in the mammograms from CC and MLO views. We estimated the detection performance of four deep learning-based AI systems using the area under the ROC curve (AUC) with a 95% confidence interval (CI). We used automatic thresholding on saliency maps from the AI systems to identify the areas of interest on the mammograms. Finally, we measured the overlap between these areas of interest and the segmented breast lesions using Dice's similarity coefficient (DSC). The detection performance of the AI systems ranged from low to moderate (AUCs from 0.525 to 0.694). The overlap between the areas of interest and the breast lesions was low for all the studied methods (median DSC from 4.2% to 38.0%). The AI system with the highest cancer detection performance (AUC = 0.694, CI 0.662–0.726) showed the lowest overlap (DSC = 4.2%) with breast lesions. The areas of interest found by saliency analysis of the AI systems showed poor overlap with breast lesions. These results suggest that AI systems with the highest performance do not solely rely on localized breast lesions for their decision-making in cancer detection; rather, they incorporate information from large image regions. This work contributes to the understanding of the role of breast lesions in cancer detection using AI.

In recent years, artificial intelligence (AI) systems have shown great potential in the analysis of screening mammograms for breast cancer detection<sup>1</sup>. One of the main difficulties for their adoption in clinical practice is the limited understanding of how these systems make their decisions<sup>2</sup>. Specifically, AI systems often involve complex computational layers with millions of parameters, which hinders the interpretation of the system's output by humans. This difficulty is often referred to as the "black box" problem of AI<sup>3,4</sup>.

To overcome this problem, explainable AI (XAI) has been developed to provide methods for understanding the decision-making of AI systems. A recent survey found that 37% of diagnostic test studies in radiology incorporate some form of XAI<sup>5</sup>. The main approaches to XAI can be classified depending on their use of visualization, semantics, or counterexamples. Visualization methods, also known as visual explanation or saliency mapping methods, are the most common form of XAI<sup>6</sup>. *Saliency mapping* is used to generate "heatmaps" that highlight the areas in medical images that most influence the prediction of the AI system<sup>7</sup>. Semantic methods, also known as textual explanation methods, provide text descriptions. Counterexamples or example-based explanation methods provide examples related to the data under analysis. A comparison of these methods of explanation is beyond the scope of this work. For a detailed review of XAI methods, we refer the reader to (Reyes et al.<sup>3</sup>) and (Van der Velden et al.<sup>6</sup>).

In this work, we use saliency analysis to compare the overlap between areas of interest from AI systems against the localization of breast lesions segmented by expert radiologists in screening mammograms. In the analysis of mammograms, the presence of suspicious regions or lesions plays a pivotal role in the decision-making process of human readers<sup>8</sup>. In turn, because AI-based systems are trained on clinical findings, it is reasonable to assume

<sup>1</sup>Escuela de Ingenierías Eléctrica Electrónica y de Telecomunicaciones, Universidad Industrial de Santander, Bucaramanga, Colombia. <sup>2</sup>Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. <sup>3</sup>Department of Radiology, Tampere University Hospital, Tampere, Finland. <sup>4</sup>Departamento de Imágenes Diagnósticas, Universidad Nacional de Colombia, Bogotá, Colombia. ✉email: otso.arponen@tuni.fi

that the lesions found in the mammograms will also play a fundamental role in AI-based analysis, which should be reflected in saliency maps<sup>9</sup>. Our hypothesis is that the areas of interest found by saliency analysis of AI systems should be associated with the localization of lesions. Thus, we expect that, for an AI-based method to perform well in breast cancer detection, saliency analysis should reveal a greater attention given to areas close to lesions. This work aims to contribute to understanding how AI systems detect malignancies by studying the role of the location of lesions in the decision-making process of these systems.

## Materials and methods

### Imaging data and lesion segmentation

To assess the detection performance of AI systems, we adopted a case–control design approach and matched cases and controls by age, screening year, and the mammographic system. We retrospectively collected mammograms from the breast cancer screening population of Tampere University Hospital. In Finland, women aged between 50 and 69 years are invited to a every two years mammographic breast cancer screening. Between 2015 and 2017, over 30,000 women were screened, and we identified 277 women with breast cancer during that period. We assessed all 277 women for the following inclusion criteria: (1) no known history of previously detected breast malignancies or previous invasive operations in the field of view (e.g., lumpectomy, mastectomy, breast implant, coiling, pacemaker), as these were hypothesized to have an impact on breast parenchyma and (2) screening-detected unilateral non-invasive or invasive cancer (i.e., no reported breast-cancer related symptoms, as they would require further management irrespective of the mammographic result). The patients who did not meet the inclusion criteria were excluded ( $N=86$ ). We identified 191 patients diagnosed with asymptomatic screening-detected, biopsy-proven cancer during the index years (2015 to 2017). All the patients who were included had been diagnosed with unilateral cancer. Corresponding healthy controls were matched by birth years, screening years, and mammographic system. The use of registered data, including mammographic images and patient history, was approved, and the need for informed consent was waived by the Research Chair of the Tampere University Hospital (permission number R18047 and R20603) in compliance with local and national regulations and laws. Because of its retrospective nature, the study was not subject to an ethics review. This retrospective study did not change either the diagnostic decisions or the management of the patients. Of the 382 women included in this study (191 cases plus 191 controls), 228 had been included in previous reports<sup>10,11</sup>. These reports concerned radiomic analysis for assessing the risk of breast cancer in *future* screening rounds, whereas in this work we are interested in breast cancer detection in the *current* screening round. There is therefore no overlap in the imaging data with previous studies.

Imaging data was retrieved from the patient flow management software (Optomed Software, Optomed Ltd, Finland). We used bilateral two-view cranio-caudal (CC) and mediolateral oblique (MLO) full-field digital mammography images (1,582 images from 382 women). Mammograms were acquired using either a MicroDose SI (Philips Healthcare [PH], the Netherlands) or a Senographe Essential (General Electric Medical Systems [GE], USA) mammography system. All images were retrieved and standardized to a resolution of 100  $\mu\text{m}/\text{pixel}$  and stored in 16-bit format. The study sample is summarized in Table 1.

For the manual delineation of breast lesions, we used software developed in-home and implemented in ImageJ<sup>12</sup>. We considered both views (CC and MLO) from the affected breast in all cases (191 women, 382 images). Lesions were manually segmented by delineating their contours on mammograms. The contour was established by consensus between a radiologist with more than 20 years of experience and a resident radiologist. It was not possible to generate segmentations for 12 images, since the lesions were only conspicuous from one of the mammographic views. As a result, 370 segmentations were considered in our saliency analysis. In this work, the manual contours drawn by the two radiologists are considered the ground truth of the breast lesions.

Characteristic	Cases (%) N = 143		Controls (%) N = 143	
Age				
< 55	30	(21)	30	(21)
55–59	40	(28)	40	(28)
60–64	53	(37)	53	(37)
> 64	20	(14)	20	(14)
Mammographic system				
Philips <sup>a</sup>	31	(22)	31	(22)
GE <sup>b</sup>	112	(78)	112	(78)
Cancer type				
DCIS	25	(17)	–	
Ductal	91	(64)	–	
Lobular	18	(13)	–	
Other	9	(6)	–	

**Table 1.** Summary of study sample. <sup>a</sup>MicroDose SI (Philips Healthcare, the Netherlands). <sup>b</sup>Senographe Essential (GE Medical Systems, USA).

## Detection performance and saliency analysis

We considered four state-of-the-art AI systems for the detection of breast cancer in screening mammography: the end-to-end deep learning architecture (End2End)<sup>13</sup>, the deep multi-view convolutional neural network (DMV-CNN)<sup>14</sup>, the globally aware multiple instance classifier (GMIC)<sup>9</sup>, and the system based on global-local activation maps (GLAM)<sup>15</sup>. These systems generate a prediction score associated with the presence of malignancies. Because we were interested in analyzing the interaction between the presence of breast lesions and saliency, all systems were set up to generate a prediction score on a per-image basis. In this study, the cancer detection task involves predicting whether there is a malignant lesion in the mammogram. Thus, the output of the systems is analyzed as a binary classification problem. The systems considered in this work are summarized in Table 2. In the table, the test population and AUCs correspond to the information reported by the respective developers of each system at the patient level. Confidence intervals are not included, since they were not reported in the original works.

Saliency analysis was performed in two steps: (1) generation of the saliency maps and (2) detection of the area of interest by saliency thresholding. For the first step, saliency maps were generated using the Grad-CAM algorithm<sup>18</sup>. Grad-CAM is a well-established method for visualization-based XAI and computes the regions in the input image that yield the highest changes in the final output of the neural network<sup>7</sup>. As a result, it generates a map that assigns a score to each input pixel according to its relevance in the decision-making process. Among the visual explanation methods for XAI, class activation mapping (CAM) methods have been the most widely used in the literature<sup>6</sup>. Gradient-weighted class activation mapping (Grad-CAM) is a generalization of CAM that can be used with any type of CNN to produce post-hoc local explanations, which are a requirement for our study.

For the sake of consistency, we normalized all scores between 0 and 1 on a per-image basis, with 1 having the highest saliency. For step (2), we identified the *area of interest* in each mammogram by selecting pixels with a saliency score above a threshold. The saliency threshold was automatically selected through an iterative process that maximizes the overlap between the area of interest and the ground truth of each lesion. This process is illustrated in Fig. 1.

The cancer detection performance of the AI systems was measured using the area under the ROC curve (AUC) with 95% confidence intervals estimated with bootstrapping. Results with  $p < 0.05$  were considered statistically significant. The AUC is a commonly used measure for assessing the performance of medical diagnosis methods due to its ability to capture the sensitivity–specificity trade-off, robustness to imbalanced data, utility in comparing different tests or models, threshold independence, and interpretability<sup>19</sup>.

The overlap between the areas of interest of each AI system and the breast lesions was measured in terms of the median Dice's similarity coefficient (DSC) with the interquartile range (IQR). DSC is a widely accepted measure for assessing the quality of segmentation methods due to its sensitivity to spatial overlap, scale invariance, ease of interpretation, and sensitivity to both false positives and false negatives. For each mammogram, the DSC estimates the overlap between a region of interest  $A$  and a lesion  $B$  as

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

DSC values range between 0 and 1, with 1 being the total overlap between  $A$  and  $B$ . A low DSC value reflects a poor co-localization between the area of interest for an AI system and the delineated breast lesion. In this work, it serves as an effective tool for objective evaluation of the overlap between the areas of interest and manually segmented breast lesions<sup>20,21</sup>.

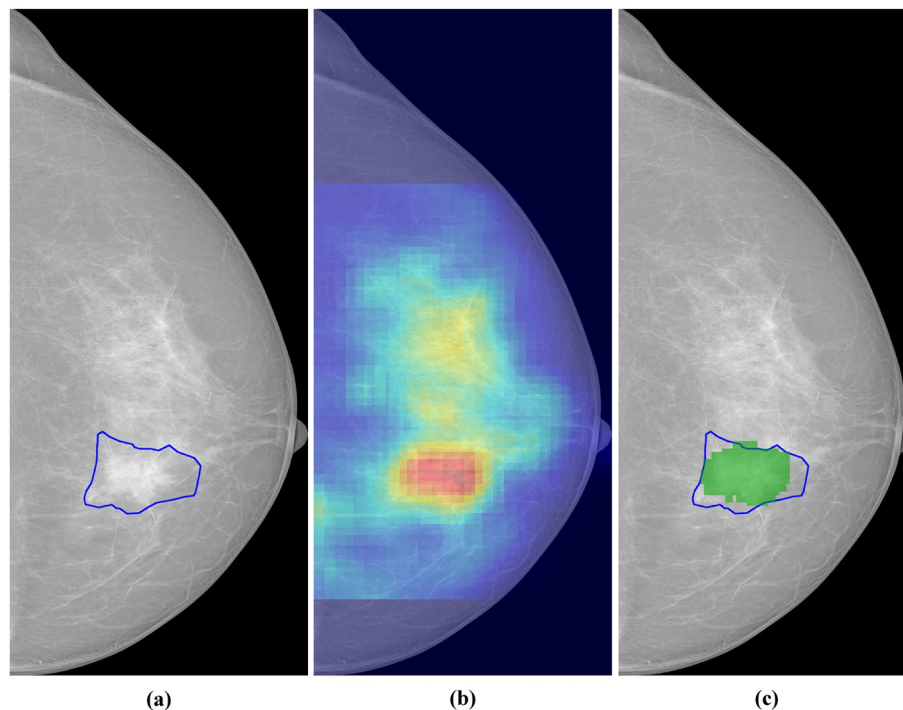
The methodology to assess the performance of the AI systems considered in this work is illustrated in Fig. 2a. The methodology for measuring the overlap between breast lesions and the area of interest of AI systems is illustrated in Fig. 2b.

## Results

In the task of cancer detection, the performance of the AI systems was low to moderate. Specifically, three of the systems yielded statistically significant performances with AUCs of 0.569, 0.572, and 0.694 for GMIC, DMV-CNN, and End2End, respectively. The performance of the GLAM method was not statistically significant. The overlap between the area of interest and the lesion in each mammogram was low for all the AI systems, with DSCs between 4.2% and 38.0%. The results are summarized in Table 3.

System	Year	Test population	AUC
GLAM <sup>15</sup>	2021	14,148 Exams from the NYU Breast Cancer Screening dataset v1 <sup>16</sup> . Mammograms: 98.6% normal, 1.2% benign, and 0.21% malignant	0.882
GMIC <sup>9</sup>	2021	14,148 Exams from the NYU Breast Cancer Screening dataset v1 <sup>16</sup> . Mammograms: 98.6% normal, 1.2% benign, and 0.21% malignant	0.930
DMV-CNN <sup>14</sup>	2020	Reader study on 720 Exams. Mammograms: 43.7% normal, 49.4% benign, and 4.3% malignant	0.886
END2END <sup>13</sup>	2019	A subset of the INbreast database <sup>17</sup> . Mammograms: 107 images from 31 women*	0.950

**Table 2.** Artificial intelligence systems for breast cancer detection. \*We include only the results reported on FFDM images.



**Figure 1.** Saliency analysis. (a) Mammogram with a manually segmented lesion. (b) Saliency map for an AI system. (c) Relevant region (in green) obtained by thresholding the saliency map in (b).

## Discussion

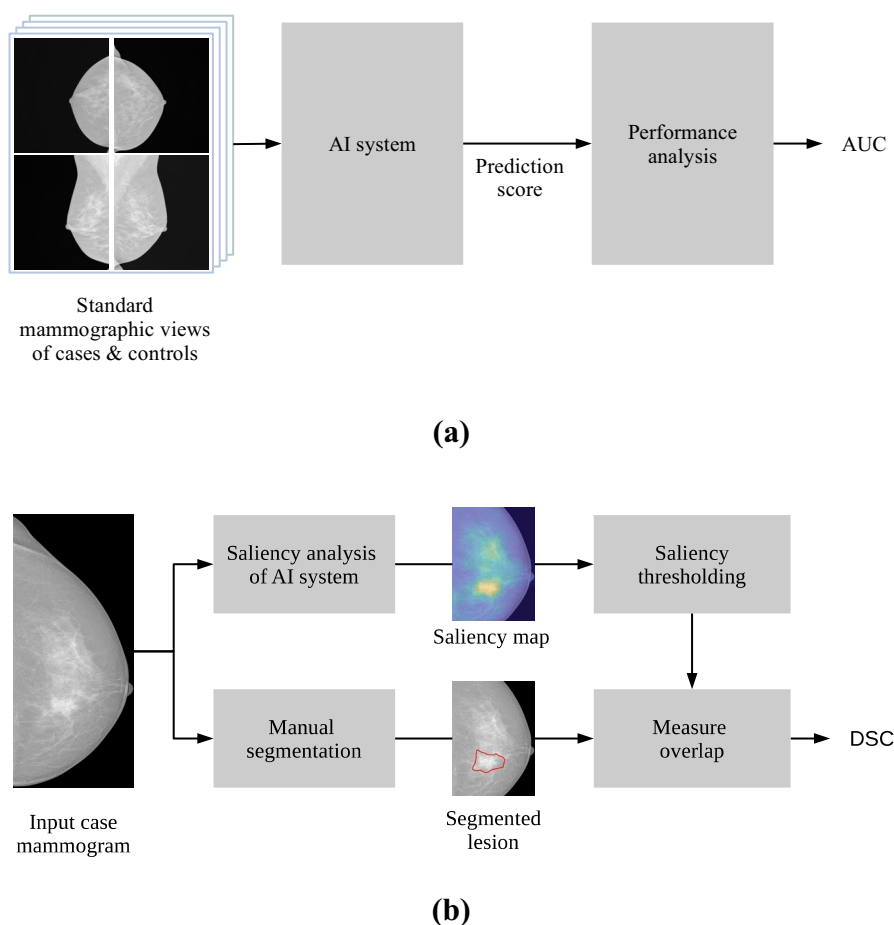
### Detection performance of AI systems

Previous research has identified external validation as one of the main difficulties for the adoption of AI systems in screening mammography<sup>22,23</sup>. Our experiments with four state-of-the-art AI systems for screening mammography showed low to moderate performance with our independent, external validation data. Comparison of the AUCs estimated by the original authors of each system (last column of Table 2) and our results (first column of Table 3) shows that the performance of all the systems decreased, with reductions in AUC ranging between 0.256 and 0.361. We argue that this reduction can be attributed to two main factors: differences in test populations and differences in experimental design. We elaborate on the implications of these differences below.

Our test population was different from the populations used in the development of all the AI systems in this study. Breast cancer epidemiology can be significantly affected by the interplay between complex factors, including the population's mean age, ethnicity, race, lifestyle, environment, socioeconomic status, and healthcare system<sup>24</sup>. To the best of our knowledge, this is the first study to assess the performance of AI systems for cancer detection in screening mammography in Finland. Our results highlight the importance of extensively testing AI systems in populations different from the ones used in the development of the systems.

There were also some important differences in our experimental design that could affect the performance of the AI systems. First, we used a case–control design matched by age and mammographic system. Age is one of the strongest risk factors for breast cancer<sup>25</sup>. In the studies where the AI systems were developed, however, age was not included in the experimental design nor accounted for in the statistical analysis. It is well known that age affects the radiological appearance of breast parenchyma<sup>26</sup>. This, in turn, could affect the performance of the AI systems: the AI systems may have relied at least partly on the age-associated changes, not the breast-related changes. Moreover, previous studies have demonstrated that differences in systems can affect the reliability of computerized mammographic analysis algorithms<sup>27,28</sup>. In fact, recent research has demonstrated the impact of technological settings on the performance of AI systems for breast cancer screening<sup>29</sup>. Finally, a previous history of breast cancer is a strong risk factor<sup>30</sup>. Changes in breast parenchyma due to previous interventions (e.g., metal clips) and treatments (e.g., radiotherapy-associated changes) can be cues for AI systems. Because we excluded symptomatic women and women with previous findings or histories of breast cancer, we believe that our experimental setting represents a more challenging scenario for the detection of breast cancer.

In addition to the aforementioned factors, previous studies have pointed to *overfitting* and *bias* as plausible explanations for the inconsistent performance of AI systems in independent test data<sup>31</sup>. A recent meta-analysis of the external validation of AI systems for screening mammography found that most studies suggest a potential diagnostic improvement when the AI systems are used together with radiologists, but warned about the persistent risk of bias<sup>22</sup>.



**Figure 2.** Experimental methodology. **(a)** Analysis of the detection performance of AI systems using the area under the ROC curve (AUC). **(b)** Saliency analysis of breast lesions using Dice's similarity coefficient (DSC).

System	AUC	95% CI	DSC (IQR)
GLAM	0.528	0.494–0.561	0.108 (0.520)
GMIC	0.571	0.535–0.606	0.291 (0.481)
DMV-CNN	0.571	0.536–0.607	0.380 (0.612)
END2END	0.688	0.655–0.719	0.042 (0.151)

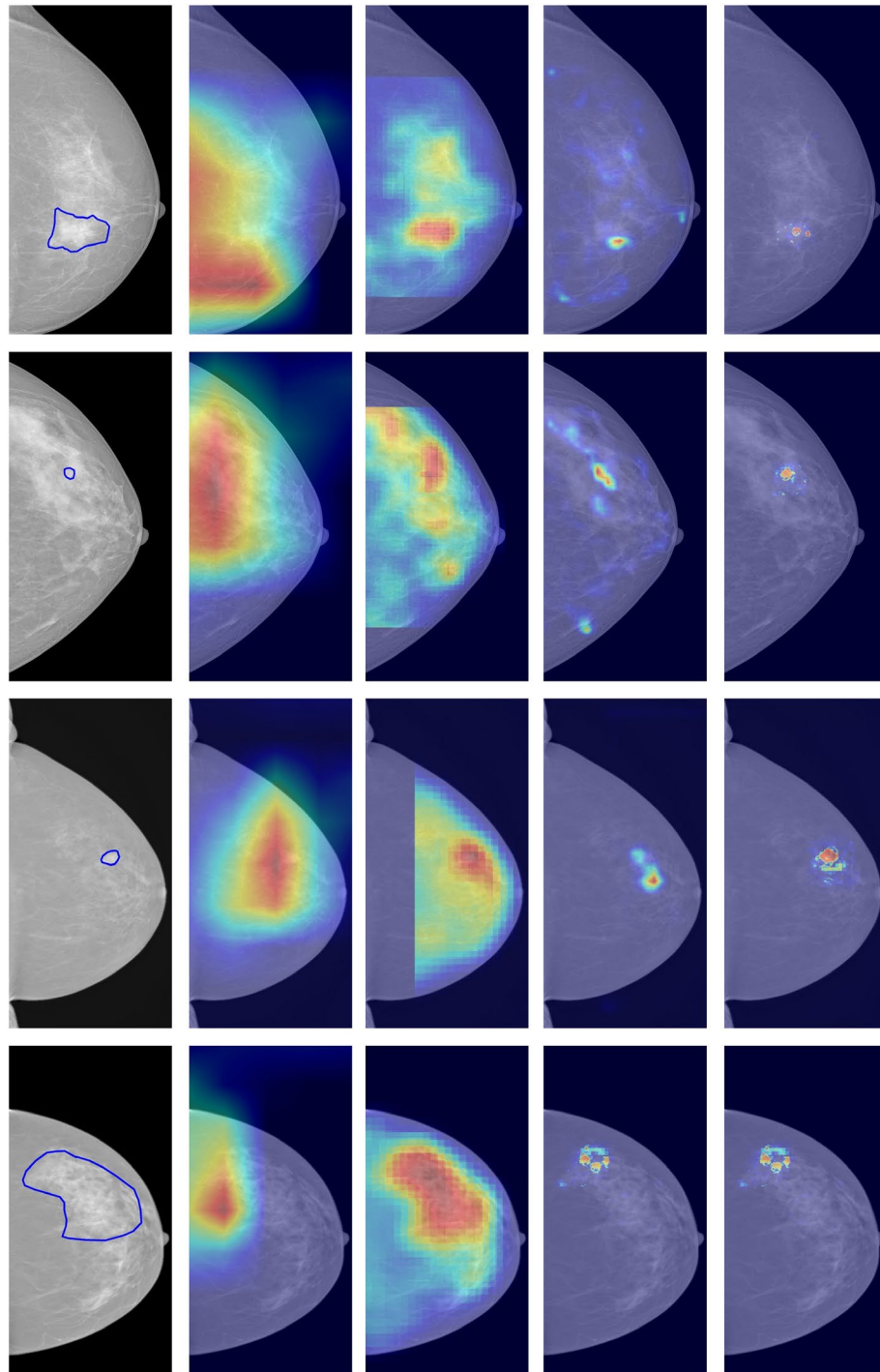
**Table 3.** Detection performance and saliency analysis.

### Relevance of breast lesions in cancer detection

In this work, we defined the *area of interest* of an AI system as the region in a mammogram with a saliency level above a threshold. For an input mammogram, this threshold was determined automatically by maximizing the overlap between the area of interest and breast lesions segmented by expert radiologists. Due to the importance of breast lesions in clinical mammographic analysis by radiologists, our hypothesis was that the area of interest of AI systems should have a high overlap with breast lesions. Our results contradict this hypothesis, however. Specifically, the DSCs showed low overlap between regions of interest and breast lesions (median DSC between 4.2% and 38.0%). In addition, the AI system with the highest performance, the End2End system with an AUC of 0.69, showed a remarkably low overlap, with a median DSC of 4.2% (IQR: 15.1%). Our results suggest that, unlike human readers, breast lesions are not as relevant to AI systems when interpreting mammograms. Specifically, the low overlap between the areas of interest and breast lesions suggests that AI systems do not rely on breast lesions as the main decision cue in diagnostics.

In recent years, the question of the interpretability of AI systems has increasingly gained attention in the machine learning community<sup>3–5,7,32</sup>. In medical imaging, XAI has been identified as one of the key factors in gaining radiologists' acceptance and, ultimately, fostering its adoption in clinical practice<sup>3</sup>. For the sake of explainability, a highly localized saliency would facilitate the understanding of what image regions or features

are more relevant for the AI system. Surprisingly, in our experiments, the highest overlap between the areas of interest and the breast lesions was observed in systems with low detection performance (AUCs between 0.52 and 0.57). As shown in the last two columns of Fig. 3, the systems with the lowest performance, GMIC and GLAM, showed highly localized saliencies. The discussion of these results should take into account the training strategy adopted for the development of the AI system. Among the methods considered in this study, GMIC and GLAM were developed to improve the “interpretability” of the AI system by focusing the analysis on localized regions of interest using labeled data. On the one hand, this helps to explain why the areas of interest of these methods



**Figure 3.** Saliency of AI systems for screening mammograms. From left to right: END2END<sup>13</sup>, DMV-CNN<sup>14</sup>, GMIC<sup>9</sup>, and GLAM<sup>15</sup>. The first two columns show the saliency maps for the best-performing systems in our study. It is clear that saliency shows a high value in a large area within each mammogram, regardless of lesion location and size.

are highly concentrated in specific spatial regions. On the other hand, the lower performance of these methods raises the question of whether the interpretability of AI systems is attained at the expense of detection performance. Our results are highly relevant for the future development of AI systems, as they show that giving a high relevance to breast lesions does not translate into a higher detection performance at the image level.

The finding regarding the relevance of large image regions for the outcome of computerized systems in the analysis of mammograms has been reported before: in breast cancer risk assessment, the extraction of high-throughput quantitative imaging biomarkers in the whole breast region, namely radiomic analysis, has consistently shown promising performance in the prediction of future breast cancer<sup>33</sup>. Based on these findings, some researchers have asked whether small changes in radiological patterns that are inconspicuous to the human eye but occupy large regions in a mammogram could play a role in the detection capabilities of computerized systems<sup>34</sup>. The fact that AI systems use information found in large image regions not circumscribed to lesions is a feasible explanation of why the joint use of AI systems with radiologists outperforms both radiologists and AI systems alone<sup>14,22,35,36</sup>. Our results suggest that successful breast cancer detection using AI systems exploits non-localized image cues not limited to breast lesions.

Our finding has great clinical significance. Recent literature has proposed that stand-alone AI algorithms could, independently or in conjunction with a radiologist, detect breast cancer or triage mammograms. Triage normal studies could be read in an adapted manner (e.g., by only one reader), and mammograms with suspicious findings could be prioritized<sup>37</sup>. AI systems that detect mammograms with findings suggestive of malignancy, albeit with limited ability to localize the tumor, would be especially beneficial for the triaging purposes of the mammograms. Such algorithms could also potentially replace one of the two readers. Nevertheless, a radiologist would still be needed to confirm the presence of the actual lesions. AI systems that localize the tumors more accurately and yet have worse performance with respect to lesion detection could be used to reduce missed diagnoses. Indeed, the results support the idea that when a developed or to-be-developed AI system is reported, the authors ought to disclose how well the system can detect mammograms with a high likelihood of breast cancer and how well it can localize the lesion.

### Limitations and future work

We identify three main limitations in our work. First, the small sample did not allow for a saliency analysis according to histopathology and tumor grading. Recent works have pointed out how the localization performance of saliency methods changes according to certain tumor-related features, such as the shape and size of lesions<sup>38</sup>. Future research should explore the performance of AI systems while considering clinical information such as breast density, tumor biology, and previous interventions and treatments. This, however, would require a substantially larger sample with annotated lesions. In this regard, we would like to highlight the importance of current efforts in the construction of large screening datasets, including clinical data and image annotations<sup>39–41</sup>.

Another limitation in this study is related to the use of saliency analysis as a means for identifying the regions that most influence the decision-making process of AI systems. Among the existing state-of-the-art methods<sup>3,6</sup>, we selected a visualization-based method, since we were interested in establishing a connection between the outcome of the AI system and specific imaging features: breast lesions. While interpreting our results, however, one should take into consideration that the interpretability of AI systems remains an open problem, and saliency does not fully explain the decision-making process of AI systems<sup>6</sup>. Few studies have measured how explainability is related to the accuracy of the system. A recent review found that, of 179 works that used XAI, only one reported measures to evaluate the outcome of the XAI<sup>5</sup>. Research on the validity of XAI methods is also scarce<sup>6</sup>. In a recent study<sup>42</sup>, the authors compared four visualization methods for pneumonia detection in chest X-rays: Grad-CAM yielded the best performance. Further investigation of saliency analysis in the context of screening mammography is warranted.

Finally, an unavoidable limitation of our work is the fact that we included a limited number of AI systems that were state-of-the-art. These systems were selected because of their good performance in previous studies and publicly available source code, which enabled the implementation of the saliency analysis. As more AI systems become available, future research is warranted to corroborate our findings.

### Conclusion

We studied the relevance of breast lesions in the decision-making of four AI systems in the mammographic analysis of breast cancer detection. For this purpose, we measured the overlap between the areas of interest in mammograms identified by saliency analysis of the AI systems and the location of breast lesions segmented by expert radiologists. The overlap between the areas of interest and lesion location was low for all four methods, whereas the best-performing methods yielded saliencies that incorporated information from large image regions in the mammogram. Our results suggest that, for the detection of breast cancer, AI systems use image cues not circumscribed to breast lesions.

### Data availability

The imaging data used in this study is not available due to restrictions on medical data. Source code and generated data are available from the corresponding author on reasonable request.

Received: 28 July 2023; Accepted: 7 November 2023

Published online: 23 November 2023

## References

1. Yoon, J. H. *et al.* Standalone ai for breast cancer detection at screening digital mammography and digital breast tomosynthesis: A systematic review and meta-analysis. *Radiology* **307**, 222639. <https://doi.org/10.1148/radiol.222639> (2023).
2. Wang, F., Kaushal, R. & Khular, D. Should health care demand interpretable artificial intelligence or accept “black box” medicine?. *Ann. Intern. Med.* **172**, 59–60. <https://doi.org/10.7326/M19-2548> (2020).
3. Reyes, M. *et al.* On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiol. Artif. Intell.* **2**, e190043. <https://doi.org/10.1148/ryai.2020190043> (2020).
4. Baselli, G., Codari, M. & Sardanelli, F. Opening the black box of machine learning in radiology: Can the proximity of annotated cases be a way?. *Eur. Radiol. Exp.* <https://doi.org/10.1186/s41747-020-00159-0> (2020).
5. Groen, A. M., Kraan, R., Amirkhan, S. F., Daams, J. G. & Maas, M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI?. *Eur. J. Radiol.* **157**, 110592. <https://doi.org/10.1016/j.erad.2022.110592> (2022).
6. van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A. & Viergever, M. A. Explainable (xai) in deep learning-based medical image analysis. *Med. Image Anal.* **79**, 102470. <https://doi.org/10.1016/j.media.2022.102470> (2022).
7. Borys, K. *et al.* Explainable AI in medical imaging: An overview for clinical practitioners—saliency-based XAI approaches. *Eur. J. Radiol.* **162**, 110787. <https://doi.org/10.1016/j.ejrad.2023.110787> (2023).
8. Sickles, E. A. *et al.* ACR-BIRADS mammography. In *ACR-BIRADS Atlas, Breast Imaging Reporting and Data System* 5th edn (American College of Radiology, 2013).
9. Shen, Y. *et al.* An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med. Image Anal.* **68**, 101908. <https://doi.org/10.1016/j.media.2020.101908> (2021).
10. Pertuz, S. *et al.* Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: A pilot study in a finnish sample. *Eur. J. Radiol.* **121**, 108710. <https://doi.org/10.1016/j.ejrad.2019.108710> (2019).
11. Padilla, A., Arponen, O., Rinta-Kiikka, I. & Pertuz, S. Image retrieval-based parenchymal analysis for breast cancer risk assessment. *Med. Phys.* **49**, 1055–1064. <https://doi.org/10.1002/mp.15378> (2022).
12. Rasband, W. ImageJ. U. S. National Institutes of Health, Bethesda, Maryland, USA, (1997–2018). Accessed on Apr. 13, 2023. <https://imagej.nih.gov/ij/>
13. Shen, L. *et al.* Deep learning to improve breast cancer detection on screening mammography. *Sci. Reports* **9**, 12495. <https://doi.org/10.1038/s41598-019-48995-4> (2019).
14. Wu, N. *et al.* Deep neural networks improve radiologists’ performance in breast cancer screening. *IEEE Transactions on Med. Imaging* **39**, 1184–1194. <https://doi.org/10.1109/TMI.2019.2945514> (2020).
15. Liu, K. *et al.* Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis. *Proc. Mach. Learn. Res.* **143**, 268–285(2021). <https://pubmed.ncbi.nlm.nih.gov/35088055/>
16. Wu, N. *et al.* The NYU breast cancer screening dataset v1.0. Tech. Rep. (New York University, 2019). <https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf>
17. Moreira, I. C. *et al.* INbreast: Toward a full-field digital mammographic database. *Acad. Radiol.* **19**, 236–248. <https://doi.org/10.1016/j.acra.2011.09.014> (2012).
18. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision* 618–626. <https://doi.org/10.1109/ICCV.2017.74> (2017).
19. Nakas, C. T., Bantis, L. E. & Gatsonis, C. A. *ROC Analysis for Classification and Prediction in Practice* (CRC Press, 2023).
20. Muller, D., Soto-Rey, I. & Kramer, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res. Notes* <https://doi.org/10.1186/s13104-022-06096-y> (2022).
21. Zou, K. H. *et al.* Statistical validation of image segmentation quality based on a spatial overlap index. *Acad. Radiol.* **11**, 178–189. [https://doi.org/10.1016/s1076-6332\(03\)00671-8](https://doi.org/10.1016/s1076-6332(03)00671-8) (2004).
22. Anderson, A. W. *et al.* Independent external validation of artificial intelligence algorithms for automated interpretation of screening mammography: A systematic review. *J. Am. Coll. Radiol.* **19**, 259–273. <https://doi.org/10.1016/j.jacr.2021.11.008> (2022).
23. Houssami, N., Kirkpatrick-Jones, G., Noguchi, N. & Lee, C. I. 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* **16**, 351–362. <https://doi.org/10.1080/17434440.2019.1610387> (2019).
24. Hortobagyi, G. N. *et al.* The global breast cancer burden: Variations in epidemiology and survival. *Clin. Breast Cancer* **6**, 391–401. <https://doi.org/10.3816/CBC.2005.N.043> (2005).
25. Lukasiewicz, S. *et al.* Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers* **13**, 4287. <https://doi.org/10.3390/cancers13174287> (2021).
26. Wolfe, J. N. Breast parenchymal patterns and their changes with age. *Radiology* **121**, 545–552. <https://doi.org/10.1148/121.3.545> (1976).
27. Pertuz, S. *et al.* Do mammographic systems affect the performance of computerized parenchymal analysis? In *International Conference of the IEEE Engineering in Medicine and Biology Society* 4863–4866. <https://doi.org/10.1109/EMBC.2019.8856948> (2019).
28. Robinson, K., Li, H., Lan, L., Schacht, D. & Giger, M. Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFD. *Med. Phys.* **46**, 2145–2156. <https://doi.org/10.1002/mp.13455> (2019).
29. de Vries, C. F. *et al.* Impact of different mammography systems on artificial intelligence performance in breast cancer screening. *Radiol. Artif. Intell.* **5**, e220146. <https://doi.org/10.1148/ryai.220146> (2023).
30. Schacht, D. V. *et al.* Importance of a personal history of breast cancer as a risk factor for the development of subsequent breast cancer: Results from screening breast mri. *Am. J. Roentgenol.* **202**, 289–292. <https://doi.org/10.2214/AJR.13.11553> (2014).
31. Wang, X. *et al.* Inconsistent performance of deep learning models on mammogram classification. *J. Am. Coll. Radiol.* **17**, 796–803. <https://doi.org/10.1016/j.jacr.2020.01.006> (2020).
32. Linaardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **23**, 18. <https://doi.org/10.3390/e23010018> (2021).
33. Gastouniotti, A., Conant, E. F. & Kontos, D. Beyond breast density: A review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast Cancer Res.* **18**, 91. <https://doi.org/10.1186/s13058-016-0755-8> (2016).
34. Hernández, A., Miranda, D. A. & Pertuz, S. An in silico study on the detectability of field cancerization through parenchymal analysis of digital mammograms. *Med. Phys.* **50**, 6379–6389. <https://doi.org/10.1002/mp.16401> (2023).
35. Schaffter, T. *et al.* Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw. Open* **3**, e200265. <https://doi.org/10.1001/jamanetworkopen.2020.0265> (2020).
36. Pacilè, S. *et al.* Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiol. Artif. Intell.* **2**, e190208. <https://doi.org/10.1148/ryai.2020190208> (2020).
37. Ng, A. Y. *et al.* Artificial intelligence as supporting reader in breast screening: A novel workflow to preserve quality and reduce workload. *J. Breast Imaging* **5**, 267–276. <https://doi.org/10.1093/jbi/wbad010> (2023).
38. Saporta, A. *et al.* Benchmarking saliency methods for chest X-ray interpretation. *Nat. Mach. Intell.* **4**, 867–878. <https://doi.org/10.1038/s42256-022-00536-x> (2022).
39. Dench, E. *et al.* Measurement challenge: Protocol for international case-control comparison of mammographic measures that predict breast cancer risk. *BMJ Open* **9**, e031041. <https://doi.org/10.1136/bmjopen-2019-031041> (2019).



40. Jeong, J. J. *et al.* The emory breast imaging dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiol. Artif. Intell.* **5**, e220047. <https://doi.org/10.1148/ryai.220047> (2023).
41. Frazer, H. M. L. *et al.* ADMANI: Annotated digital mammograms and associated non-image datasets. *Radiol. Artif. Intell.* **5**, e220072. <https://doi.org/10.1148/ryai.220072> (2023).
42. Arun, N. *et al.* Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Artif. Intell.* **3**, 267. <https://doi.org/10.1148/ryai.2021200267> (2021).

## Acknowledgements

G. Africano was funded by the project “Software de análisis parenquimatoso de imágenes mamográficas para la evaluación de riesgo de cáncer de seno” (MINCIENCIAS, 110284467139). S. Pertuz was partially funded by the project “Estudio piloto para el desarrollo y evaluación de descriptores cuantitativos de imágenes de ultrasonido transvaginal para predicción de parto prematuro” (Universidad Industrial de Santander, VIE3947). D. Ortega was funded by the grant “Convocatoria 907 Jóvenes Investigadores e Innovadores en el marco de la reactivación económica 2021.” The authors also thank Ing. Marly Gallo for her early contributions to the development of the lesion annotation tool.

## Author contributions

All authors contributed to the planning of this manuscript. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to O.A.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023