European Journal of Radiology 175 (2024) 111457

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Review

Artificial Intelligence for breast cancer detection: Technology, challenges, and prospects

Oliver Díaz [a,b], Alejandro Rodríguez-Ruíz [c], Ioannis Sechopoulos [d,e,f,*]

[a] Artificial Intelligence in Medicine Laboratory, Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain
[b] Computer Vision Center, Barcelona, Spain
[c] ScreenPoint Medical, Nijmegen, the Netherlands
[d] Department of Medical Imaging, Radboud University Medical Center, Nijmegen, the Netherlands
[e] Dutch Expert Centre for Screening (LRCB), Nijmegen, the Netherlands
[f] Technical Medicine Center, University of Twente, Enschede, the Netherlands

ARTICLE INFO

Keywords:
Breast cancer
Breast imaging
Artificial intelligence
Mammography
Screening

ABSTRACT

Purpose: This review provides an overview of the current state of artificial intelligence (AI) technology for automated detection of breast cancer in digital mammography (DM) and digital breast tomosynthesis (DBT). It aims to discuss the technology, available AI systems, and the challenges faced by AI in breast cancer screening.
Methods: The review examines the development of AI technology in breast cancer detection, focusing on deep learning (DL) techniques and their differences from traditional computer-aided detection (CAD) systems. It discusses data pre-processing, learning paradigms, and the need for independent validation approaches.
Results: DL-based AI systems have shown significant improvements in breast cancer detection. They have the potential to enhance screening outcomes, reduce false negatives and positives, and detect subtle abnormalities missed by human observers. However, challenges like the lack of standardised datasets, potential bias in training data, and regulatory approval hinder their widespread adoption.
Conclusions: AI technology has the potential to improve breast cancer screening by increasing accuracy and reducing radiologist workload. DL-based AI systems show promise in enhancing detection performance and eliminating variability among observers. Standardised guidelines and trustworthy AI practices are necessary to ensure fairness, traceability, and robustness. Further research and validation are needed to establish clinical trust in AI. Collaboration between researchers, clinicians, and regulatory bodies is crucial to address challenges and promote AI implementation in breast cancer screening.

1. Introduction

Automated detection of breast cancer in mammography using computer algorithms is a decades-old topic of research, development, and clinical use [1]. These research and development efforts have been driven by the desire to assist radiologists during the challenging task of detecting early signs of breast cancer in a mammogram. Currently, whether using digital mammography (DM) or digital breast tomosynthesis (DBT), breast cancer screening still relies only on radiologist (for the most part) assessments, who, even at their best performance, still overlook some visible cancer lesions (false negative assessments) and decide to recall healthy women for further assessment (false positive assessments) [2]. In addition, screening programs come at an expensive workload, since the vast majority of exams are normal [3]. This is compounded in screening programs, where it is common for all exams are read by at least two radiologists, as is commonly the case in Europe. Therefore, screening programs still have room for improvement, and researchers are continuously looking into technological breakthroughs that can improve screening outcomes by increasing accuracy and/or reducing radiologist workload. Computer-aided detection (CAD) systems for mammography using traditional machine learning (ML) artificial intelligence (AI) techniques have been available since the 1990s [4]. These ML-based CAD systems automatically analyse the images and display suspicious areas on the mammogram. Radiologists used these systems as a second-look aid, to reduce the chances of overlooking errors. However, their high false positive rate was one of the reasons that led to a poor acceptance of these systems in European screening programs [5]. More recently, new systems have been developed using deep

* Corresponding author at: Department of Medical Imaging, Radboud University Medical Center, P.O. Box 9101 (766), 6500, HB, Nijmegen, the Netherlands.
E-mail addresses: oliver.diaz@ub.edu (O. Díaz), ioannis.sechopoulos@radboudumc.nl (I. Sechopoulos).

https://doi.org/10.1016/j.ejrad.2024.111457
Received 27 March 2024; Accepted 8 April 2024
Available online 16 April 2024

learning (DL) AI techniques, which have been proven to significantly enhance the breast cancer detection performance of radiologists [6,7,8]. Numerous articles have been published and dozens of conference talks have been presented, preaching how DL-based AI systems (hereinafter simply AI systems) are the technological breakthrough that will finally improve breast cancer screening programs, whether using DM or DBT [1,9]. Recent AI systems have been praised to be "as good as radiologists reading mammograms", demonstrated to "help radiologists be more accurate detecting breast cancer", as well as to be a solution that can "(partially) replace radiologists reading screening mammograms". However, as with any promised technological improvement, more thorough research is required to ensure their successful implementation.

This review article aims to provide an overview on the use of AI technology for automated detection of breast cancer in DM and DBT, including a discussion on relevant issues such as technology description, currently available AI systems for breast cancer screening, and the challenges and open questions still faced by the numerous AI systems in the market and the technology in general.

2. Breast AI technology

2.1. AI and DL: Concepts and differences with traditional CAD

The role of AI in medical imaging has been expanded enormously in the last years [10] especially in the field of breast imaging [1]. This has been possible due to the rapid increase in affordable computer power (and availability of cloud computing resources), which allows for billions of operations per second, the availability of digital data, and the capacity to store this digital information inexpensively.

AI includes a wide ecosystem of techniques within the computer science field where machines are programmed to simulate human intelligence, including the capacity to learn. AI covers many subfields, such as ML, and DL, among others (see Fig. 1). AI technology is found in countless devices, such as computers, mobile phones, smartwatches, cars, etc. Its applications are very wide, from management of junk emails, robotics, and voice processing, to clinical decision support systems.

The first computer aided detection/diagnosis (CADe/CADx) tools appeared in the 1990s and used traditional ML strategies [4]. In these, known morphological, intensity, or texture image features were extracted and processed in a specific model (e.g., decision trees, support vector machines, etc.) in order to find patterns in the data that allowed performing a given task (prediction, detection, etc.). This type of approach has certain limitations, such as the need of a large amount of hand-crafted, structured training data or human supervision, since most of the learning is generally done in a supervised manner.

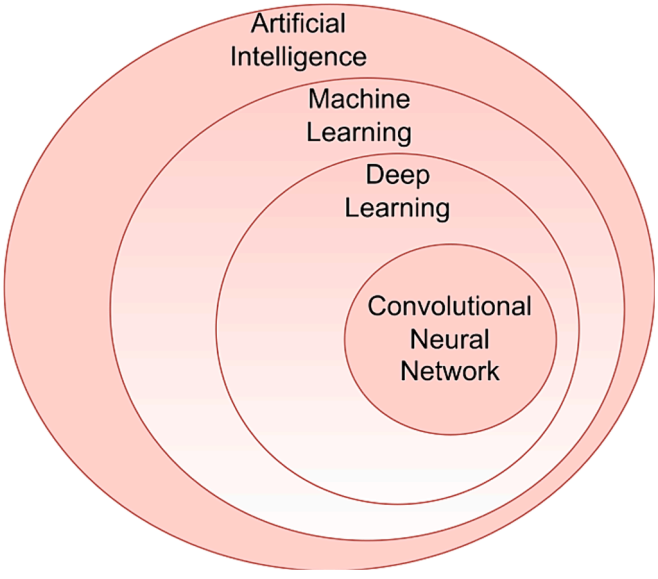

Fig. 1. Diagram showing the artificial intelligence ecosystem with several of its subfields.

DL-based CADe/CADx tools belong to a more recent and popular subtype of ML technique that uses one or more neural network (NN) architectures (see Fig. 2), similar to the neuron connections within the human brain, to create models capable of making accurate data-driven decisions. It differs from traditional ML since the features are not manually selected but learned and tuned at training time through a complex optimisation process better known as backpropagation [11].

DL algorithms come in many forms depending on the architecture employed, these include deep NN (DNN), recurrent NN (RNN), deep belief networks (DBN), and convolutional NN (CNN). CNNs are the most commonlyused algorithms for image segmentation and classification, although a more recent architecture called Transformer [12] is significantly improving upon the results of current state-of-the-art CNNs [13]. CNNs gained great interest after their excellent results in the ImageNet image (regular photos) classification competition in 2012. In this competition, the AlexNet CNN [14] outperformed all other techniques by a large margin. Since then, NNs have continued to progress and are now used for a variety of problems and types of datasets. This is because DL models have the ability to learn relevant features directly from datasets, enabling more effective analysis and interpretation. Recently-developed DNNs contain more neurons than previous networks and can have more complex ways of connecting these layers of neurons, resulting in more complex functions that can represent robust features for medical segmentation [15].

DL has had huge success in many research areas of medical image analysis, although its clinical implementation has been limited. For instance, there are reports of large performance improvements in breast cancer CADx and CADe systems by analysing and integrating diagnostic information [16,17]. In recent years, Rodríguez-Ruíz et al. [18] showed that NNs can achieve the detection rates of an average radiologist. More recently, Lång et al. reported on the first randomized screening trial, in which DL not only substantially reduced the radiologist workload, but also improved the sensitivity for cancer detection with no loss in specificity [8,19]. In addition, NNs allow for the development of classification models to distinguish between malignant or benign lesions. This has the added benefit that the automated classification performed by NNs eliminates the variability that can exist among observers.

Many AI examples can be found in the field of breast cancer detection, mostly DL based, using such diverse imaging technologies as input signals such as DM [20,21,22], DBT [23,24], and breast MRI [25]. Since 3D and 4D medical imaging are becoming routine in clinical practice, and with physiological and functional imaging capabilities increasing, medical imaging data is increasing in size and complexity. Therefore, AI

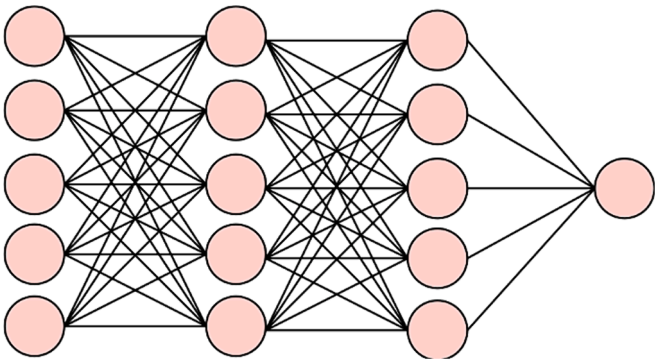

Fig. 2. Diagram of a neural network. Each circle is a node (neuron or perceptron) that combines the data from prior layer (of nodes) with a set of coefficients (or weights).

will play a key role in processing such vast amounts of data. In fact, although the majority of the current DL-based breast cancer detection tools described in the literature focuses on a single imaging modality and task, the current trend of AI is to combine multi-modal input data together with nonimaging data to arrive at more accurate outcomes.

2.2. Training AI systems: The importance of having good data

Currently, the value of data is such that it is sometimes referred to as the oil of the 21st century [26]. Similar to the need to refine oil for it to be useful, the raw data needs to be processed so it can be transformed into usable information, knowledge, and eventually, wisdom [27]. In fact, during the life cycle of data mining projects, data scientists have reported that up to 60 % of their time is spent in data cleaning and processing [28].

The quality of the data is of utmost importance since AI algorithms will use such data to extract patterns and features that will be eventually used to provide outcomes. Thus data pre-processing strategies, which cover acquisition, data curation, annotation, and data storage (while maintaining patient privacy) is key. Pre-processing is also important in the medical imaging domain for the optimal performance of AI algorithms [29]. Also, a combination of imaging and non-imaging data is crucial for AI models to provide more clinically relevant results. Therefore, it is recommended to spend sufficient time and care in this stage to maximise the AI-powered results. However, a good pre-processing step does not guarantee success in the AI results, since model development and training is also critical. However, poor data quality cannot be simply overcome by more sophisticated models.

Several types of learning paradigms are used by AI algorithms (see Fig. 3). The most common are supervised, unsupervised, semi-supervised (a mixture of supervised and unsupervised), and reinforcement learning, although other types of learning also exist. Once trained, AI models assign a weight to each node (e.g., image feature) to highlight its importance in the final task of the network (segmentation, detection, etc.).

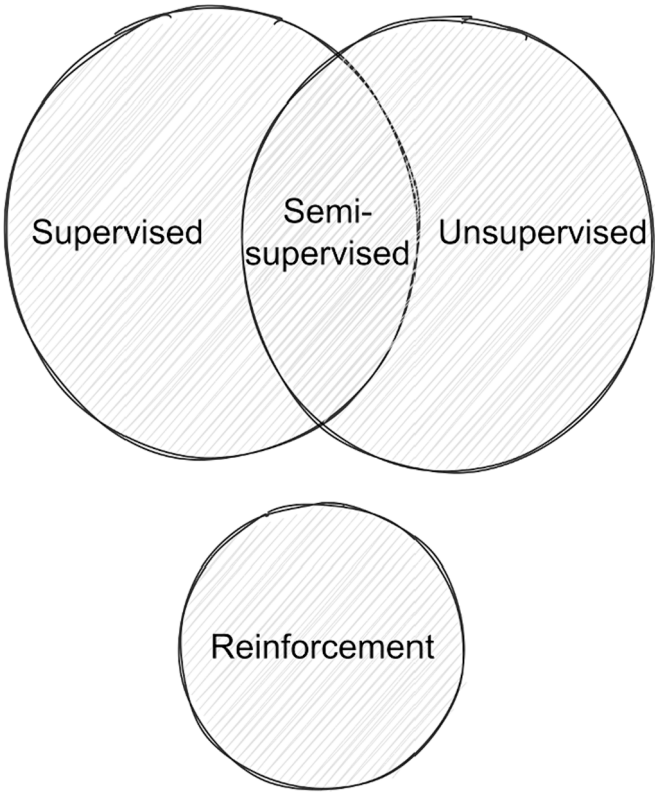

Fig. 3. Most common learning paradigms used to train AI models.

In supervised learning, the data is fully annotated. Therefore, there is a priori knowledge of the ground truth associated with each data item (i.e., label). For example, in the breast imaging domain, a mammogram can have a label of "cancer"/"no cancer" for the detection task, "benign"/"malignant" for classification purposes, or a bounding box around the edge of a lesion in case of segmentation/detection. Unsupervised learning models use untagged images where the algorithm tries to group data according to common patterns in the data. Semisupervised learning combines both supervised and unsupervised learning features. In other words, the AI model learns from labeled and unlabeled data. Reinforced learning uses an agent (e.g., cancer detector) to learn in an interactive environment (e.g., mammogram) by trial and error. This agent uses the feedback from its own actions and experiences to learn optimal behaviour.

During the development of AI algorithms, the data is typically split into training, validation, and test datasets. Since many datasets are not large enough, several strategies can be used to increase the number of samples (i.e., data augmentation). Original data can be augmented by applying certain operations (e.g., rotation, scaling) that should not affect the expected decision from the algorithm. Also, synthetic data generated by adversarial networks are also used in cancer imaging for data augmentation, or even to populate the less frequent classes (e.g., rare cancer types, low or high dense breasts) [30,31,32,33,34,35]. Finally, transfer learning strategies are also commonplace, in which networks already trained with other data (with their corresponding weights) are used to train a new model with new, but less, data in a similar or different domain. This fine tuning typically requires a reduced training time since small adjustments are needed to achieve the desired output or performance for the new type of data.

2.3. Validation of AI systems: independent, comprehensive, and generalisable

AI tools, especially DL-based, are often perceived as black boxes, or grey in the best of cases, by many healthcare actors, where results seem magically generated. Although AI is generally accepted with excitement by the healthcare community [36,37,38], the lack of insight into this magical world of AI frequently reduces the confidence in its results and hinders its clinical implementation [39]. This also includes CE-marked AI products for clinical radiology, where a recent study found that 64 % of the AI algorithms reviewed (n = 100) had no scientificallyproven evidence of their clinical efficacy [40].

In order to improve AI acceptance, the European Commission, through the European High-Level Expert Group on Artificial Intelligence (AI HLEG), published in 2020 the assessment list for (general) trustworthy AI (ALTAI) [41]. However, these recommendations refer to AI in general and do not address the specific risks and challenges found in healthcare.

Therefore, there are several international initiatives to provide checklists and guidelines for researchers and clinicians to improve the trustworthiness of AI in medical imaging: TRIPOD-AI/PROBAST-AI [42], CLAIM [43], MINIMAR [42] CONSORT-AI [44], CLEAR citekocak2023checklist, metrics reloaded [45] or FUTURE-AI [46].

For example, the FUTURE-AI guidelines (https://future-ai.eu/) provide six essential principles (Fairness, Universality, Traceability, Usability, Robustness, and Explainability) to increase the clinical trust and adoption of AI technology in medical imaging, as described below.

Fairness. AI algorithms should maintain the same performance when applied to similar individuals (individual fairness) and across subgroups of individuals, including under-represented groups (group fairness). For example, the less frequently-encountered breast density groups in datasets could be mitigated through data augmentation strategies [35,34].

Universality. This principle recommends the definition and application of (technical, clinical, ethical, and regulatory) standards during

AI development, evaluation, and deployment. This will increase the interoperability and applicability of AI tools across clinical centres.

Traceability. Medical AI algorithms should be developed with mechanisms for documenting and monitoring the development process, as well as be paired with methods to continuously or periodically monitor their functioning in the clinical environment. For example, monitoring tools to identify any discrepancy or drift from its expected behaviour, both in terms of accuracy and interaction with end users, and to plan maintenance interventions for nurturing the AI models.

Usability. Medical AI tools should be usable, acceptable, and deployable for the real-world end users (i.e., physicians, radiologists, and other end users). For example, studies evaluating the usability of the algorithm in question should include all actors expected to use and interact with the system.

Robustness. This principle refers to maintaining AI's performance and accuracy when it is applied under the highly variable conditions that could be encountered in the real world, outside the controlled environment of the laboratory where the algorithm was built and initially tested. For example, development and validation of AI products should include the use of multi-centre and multivendor datasets.

Explainability. Medical AI algorithms should be able to provide meaningful and actionable explanations of their predictions to the end users. Explainability provides insight into the algorithmic mechanisms behind the AI decision-making processes.

Two key components to improve the clinical implementation of AI models are the development of independent and generalisable AI models, through training with good data and the production of AI-powered transparent or explainable results.

On the one hand, the dataset used during training, testing, and validation should be fair and free of bias (e.g., based on ethnicity, country, manufacturer, etc.) to avoid potential unfair treatment of certain groups. This is why multicentre and multivendor datasets are highly recommended, resulting in a more generalisable AI model [47]. Despite this data requirement, privacy-related issues sometimes hinder the access of such broadlyencompassing data. Problems derived from ethical committees, (pseudo-)anonymisation, encrypted data transfer, and secure storage can make it difficult to build a good dataset for AI training and validation. However, novel techniques, such as federated or swarm learning, allow for the training of AI models within a clinical site without data sharing [48]. With federated learning, AI models are trained locally and the resulting model's coefficients (i.e., weights) are sent outside the clinical site where they are combined/aggregated with models trained in other centres. This way, sensitive patient information never leaves the secure network of the clinical site where the data resides.

On the other hand, AI-powered results should provide predictions that can be understood by both AI developers and clinical users. In order to achieve this, explainable AI (XAI) was born [49,50]. In general, AI models work with probabilities that are later thresholded to provide a decision (e.g., detection). Such a result may be difficult to believe if it is not supported or justified by further information. Interpretability models are vital to describe the AI results to healthcare professionals (e.g., radiologists), improve their trust in the algorithms, thus these could be incorporated into the decision-making process. Several strategies exist to explain AI-derived results, but the most common ones are based on visual explanations [51]. For example, attribution maps, such as Grad-CAM [52], use heat maps to highlight the pixels in the image with higher weights in the final decision and therefore considered most important by the AI model (Fig. 4).

3. How can I use breast AI in screening?

2.4. From (partially) replacing radiologists to being used for concurrent support

AI systems can be used in different ways in breast cancer screening with DM or DBT. The different possibilities are driven by local needs and/or preferences and requirements of each screening program. Based on scientific literature, Table 1 presents the most important approaches that have been proposed for the implementation of breast AI in screening.

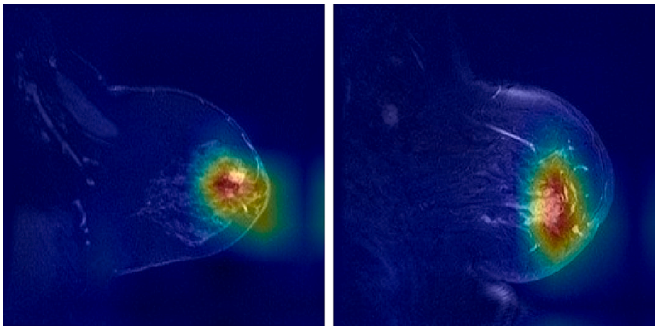

Fig. 4. Breast MRI with a superimposed heatmap. The hottest region represents the pixels with higher influence used by an AI model to arrive at a prediction. Courtesy of Ms Smriti Joshi from University of Barcelona.

As expected, the potential impact in screening is different depending on the specific implementation of breast imaging AI (Table 1). For example, some approaches can dramatically reduce screening workload without reducing sensitivity, while others can increase sensitivity but at the cost of also increasing false positives. It is yet to be demonstrated (with real-life evidence) if introducing breast AI in screening could bring a three-fold improvement in workload, sensitivity, and specificity, although Lång et al. seems to be getting us close [8,19].

It should be noted that the above-mentioned uses are not independent from each other, and therefore could be combined. Although the field of breast imaging AI has one of the largest bodies of evidence within radiological AI, the degree of evidence is still limited. Most of the studies are based on retrospective analyses of data, only the minority of commercially available AI systems are investigated, and often the studies are performed with limited data, acquired in a single site, or without enough heterogeneity in the data to be representative of screening programs worldwide. Other issues, such as the enrichment of datasets, especially in screening, the lack of consideration of how radiologists will change their behaviour when AI has a role in the interpretation, and the limited clinical relevance of results (e.g., indolence or aggressiveness of screen-detected cancers), are also common limitations in the current literature.

Ultimately, the possible use of breast imaging AI technology is also constrained by the characteristics of each AI system, such as its actual intended use following regulatory clearance and its features and performance. For example, an AI system that has been cleared only for concurrent decision support for radiologists during mammogram interpretation is not regulatory approved to be used as a stand-alone reader of mammograms in screening, even if part of a double reading scenario.

3. Challenges and prospects

3.1. What AI system should I use? How do I know if I can use this product in my hospital?

With currently over half a dozen breast imaging AI systems available in the market, the user may often be challenged with the question of which AI system to implement in their clinical setting.

Although some commercial AI systems are more accurate than others as illustrated in the work by Salim et al. [67], other elements are equally important, such as the intended use of each system, the deployment possibilities, and the actual performance for the user's specific population and equipment.

It is recommended for users to perform periodic evaluations or audits

Table 1
Overview of different strategies proposed for implementation of AI system in breast cancer screening with mammography or tomosynthesis and their measured impact.

Strategy: Use AI as concurrent decision support				
Publication	Study Type	Modality	Impact on Screening or Human Performance	Impact on Workload
Pacile et al. [53]	Reader study	DM	Improved radiologist AUC from 0.769 without AI to 0.797 when reading with AI (P=.035).	9–14% longer than reading time per case when using AI (P < .001).
Conant et al. [23]	Reader Study	DBT	Improved radiologist AUC from 0.80 without AI to 0.895 when reading with AI support (P<.01).	53% faster reading time per case when using AI (P < .01).
Rodriguez-Ruiz et al. [54]	Reader Study	DM	Improved radiologist AUC from 0.87 without AI to 0.89 when reading with AI support (P =.002).	Similar reading time per case when using AI (P =.15).
Van Winkel et al. [55]	Reader Study	DBT	Improved radiologist AUC from 0.83 without AI to 0.86 when reading with AI support (P <.001).	12% faster reading time per case when using AI (P =.003).
Pinto et al. [56]	Reader Study	DBT	Improved radiologist AUC from 0.85 without AI to 0.88 when reading with AI support (P =.01).	Similar reading time per case when using AI (P =.35).
Kim et al. [57]	Reader Study	DM	Improved radiologist AUC from 0.81 without AI to 0.88 when reading with AI support (P <.0001).	Not reported.
Strategy: Use AI as an independent stand-alone 2nd reader of screening				
Publication	Study Type	Modality	Impact in Screening or Human Performance	Impact on Workload
Dembrower et al. [58]	Prospective Paired Study (ScreenTrustCAD)	DM	4% higher CDR (P = 0.017) and 4% lower recall rate (P < 0.05).	50% fewer screening readings needed.
Larsen et al. [59]	Retrospective evaluation of a large screening sample	DM	Same CDR and 16% lower recall rate.	50% fewer screening readings needed.
Sharma et al. [60]	Retrospective evaluation of a large screening sample	DM	Non-inferior CDR and recall rate.	30-45% reduced screening workload.
Leibig et al. [61]	Retrospective evaluation of a large screening sample	DM	Lower CDR and higher recall rate.	50% fewer screening readings needed.
Strategy: Use AI as triage tool, low risk exams are single read and high-risk exams are double read				
Publication	Study Type	Modality	Impact in Screening or Human Performance	Impact on Workload
Lång et al. [8]	Randomized Control Trial (MASAI)	DM	Trend for +20% CDR (P = 0.052) and same recall rate.	44% fewer screening readings needed.
Larsen et al. [59]	Retrospective evaluation of a large screening sample	DM	Same CDR and 9% lower recall rate.	35% fewer screening readings needed.
Strategy: Use AI as triage tool, low risk exams are automatically labelled as normal and high-risk exams are double read				
Publication	Study Type	Modality	Impact in Screening or Human Performance	Impact on Workload
Lång et al. [62]	Retrospective evaluation of a large screening sample	DM	Same sensitivity and recall rate.	19% fewer screening readings needed.
Raya-Povedano et al. [63]	Retrospective evaluation of a large screening sample	DM	Same sensitivity and 17% lower recall rate (P < 0.001).	71% fewer screening readings needed.
Raya-Povedano et al. [63]	Retrospective evaluation of a large screening sample	DBT	Same sensitivity and 17% lower recall rate (P < 0.001).	72% fewer screening readings needed.
Larsen et al. [59]	Retrospective evaluation of a large screening sample	DM	Same CDR and 19% lower recall rate.	50% fewer screening readings needed.
Lauritzen et al. [64]	Retrospective evaluation of a large screening sample	DM	Same sensitivity and 19% lower recall rate.	63% fewer screening readings needed.
Dembrower et al. [65]	Retrospective evaluation of a large screening sample	DM	Potential additional CDR of 71 per 1000 examinations.	60% fewer screening readings needed.
Shoshan et al. [66]	Retrospective evaluation of a large screening sample	DBT	Noninferior sensitivity (P =.002), and 25% lower recall rate (P = .002).	40% fewer screening readings needed.

CDR = Cancer Detection Rate. DM = Digital Mammography. DBT = Digital Breast Tomosynthesis.

of their AI algorithms before and during implementation, to monitor that the system errors are not threatening the safety and effectiveness of the medical service. It is encouraged that AI developers provide accessible tools to the user to run a cohort of cases through their algorithms, to monitor performance, which promotes transparency and may help the user decide on whether to introduce a product into their clinical environment or not.

Well-established guidelines and checklists accepted by the medical imaging community will be pivotal to improve the trustworthiness of future AI algorithms by radiologists. Traceability measures should be put in place together with mechanisms for documenting and monitoring the development and functioning of the AI tools in the clinical environment. The availability of information regarding the datasets used to train and validate an algorithm, for example, would increase the transparency and quality assurance of AI to help make decisions regarding the acquisition of a given product. Obviously, internal

validation should be performed using local data to detect error or biases (e.g., images trained with systems from a single vendor) that could have been present during training of the algorithm.

3.2. What are the next technological improvements that are needed to make AI even closer to acting as a radiologist in breast cancer screening?

Most of the AI-based algorithms for breast cancer detection described in the literature employ a single DM or DBT image (or even patches of images) from a single time point to make predictions. However, in real clinical scenarios, radiologists also consider other sources of radiological information that are extracted from prior images, contralateral images, or even other data from the patient medical record. Future AI-based decision-making algorithms should take into account all relevant multi-source data. This fact will require further effort to compile such a large amount of data and information, but it is key to allow software tools to make more accurate predictions.

This requires more clinical data being available for training the algorithms and patient privacy-preserving strategies, such as federated learning or synthetic data generation, will play an important role in the future development of AI algorithms.

3.3. How to keep up to date with the evolving field of AI?

Although there is a plethora of information on AI developments in each radiology journal issue and conference, there is a gap between the information available in peer-reviewed publications, usually based on experiments performed under strict controls, and real-life use. The true impact of AI under real-world clinical conditions is harder to gauge, quantify, compile, interpret, and communicate. Methods to aggregate and publicise performance benchmarks of AI in actual clinical use, of the type similar to that achieved by the Breast Cancer Surveillance Consortium [68] for regular screening and diagnostic work, could go a long way in providing the community with information on the actual impact of these new algorithms.

3.4. Do we have to wait for the results of prospective clinical trials?

So far, most of the evidence evaluating the impact of AI systems in breast cancer screening is based on studies simulating the use of AI in retrospectively collected data cohorts, or in laboratory environments studying AIhuman interaction. The paucity of evidence based on reallife AI use is one of the biggest barriers to the widespread adoption of AI in breast cancer screening. Nevertheless, the first results from a prospective trial have been recently published. MASAI [8] was a randomised control trial evaluating the impact of using AI to triage DM exams that require single reading from those needing double reading. Latest results presented at ECR 2024 showed a 28 % higher CDR ($P = 0.002$) and similar recall rate (2.1 % (2.02.2) vs 1.9 % (1.8–2.1)). The impact on interval cancer detection rates will also be evaluated and reported after all the women are followed-up. The ScreenTrustCAD [58] study also gave some indication of what can be expected when AI acts as an independent reader in a double reading setting with DM. Results showed 4 % higher CDR ($P = 0.017$), 21 % higher arbitration rate (relative proportion 1·21 (1·18–1·24)) and 4 % lower recall rate (relative proportion 0·96 (0·94–0·97)). This was achieved reducing by 50 % the number of screen readings.

The results from these trials validate the promising results seen in retrospective evaluations of AI in breast cancer screening.

According to clinicaltrials.gov, there is another ongoing prospective clinical trial aiming to evaluate the impact of AI when implementing it in real-life screening with DM and DBT. The AITIC trial (Artificial Intelligence in Breast Cancer Screening Programs in Cordoba) is evaluating the hypothesis that AI can be used in a partially autonomous strategy, removing the need for humans to read up to 70 % of the most likely normal screening exams (either DM or DBT), without reducing the cancer detection rate. Preliminary results presented at ECR 2024 after including 24,000 women demonstrate that this strategy is safe and effective and could lead to a future where lowrisk screening exams are automatically labeled as normal. In their study, CDR increased by 1.0/1000 from 6.2/1000 reading 100 % of screening exams without AI to 7.2/1000 when reading only the 35.6 % most suspicious exams with AI support, without a decrease in the recall positive predictive value.

An open question is whether countries and screening programs will leverage prospective evidence from other regions, or each will want to conduct their own trial to evaluate the use of AI according to their local needs, thus delaying the introduction of the potential benefits of AI in screening. On the other hand, many hospitals are already implementing AI in institution-based screening, highlighting the fact that local clinical needs are one of the main drivers to determine how and when to use AI for breast cancer screening.

4. Conclusions

The use of AI technology for automated detection of breast cancer in mammography has shown significant progress and holds great potential for improving screening programs. Traditional CAD systems using machine learning techniques have been available for decades, but their high false positive rate limited their acceptance in screening programs. However, recent advancements in DL-based AI techniques have demonstrated improved performance in breast cancer detection, surpassing the capabilities of radiologists in some cases.

The application of DL-based AI systems in breast cancer screening, whether using DM or DBT, has been widely researched and discussed. These AI systems have been praised for their ability to assist radiologists in detecting breast cancer more accurately and reducing false negative and false positive assessments. They have even been suggested as a potential replacement for radiologists in reading screening mammograms, although further research is needed to ensure successful implementation.

Training AI systems requires high-quality data, and data preprocessing is crucial for optimal performance. Supervised, unsupervised, semi-supervised, and reinforcement learning paradigms are used, with data augmentation techniques and transfer learning strategies being employed to overcome limited data availability. Validation of AI systems is essential to ensure their trustworthiness and clinical efficacy, and international initiatives are providing guidelines and checklists for researchers and clinicians to improve transparency and reliability.

Adhering to these initiatives will contribute to the increased acceptance and adoption of AI tools in clinical practice. While AI shows great promise in improving breast cancer detection, further research, collaboration, and validation efforts are needed to ensure its successful integration into routine screening programs.

CRediT authorship contribution statement

Oliver Díaz: Conceptualization, Data curation, Writing – review & editing, Writing – original draft. **Alejandro Rodríguez-Ruíz:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Ioannis Sechopoulos:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

O.D. has received funding from the European Union's Horizon Europe and Horizon 2020 research and innovation programme under

grant agreement No 101057699 (RadioVal) and No 952103 (EuCanImage), respectively.

References

[1] I. Sechopoulos, J. Teuwen, R. Mann, Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: state of the art, in: Seminars in Cancer Biology, vol. 72, Elsevier, 2021, pp. 214–225.

[2] H.J. Schunemann, D. Lerda, C. Quinn, M. Follmann, P. Alonso-Coello, P.G. Rossi, A. Lebeau, L. Nystrom, M. Broeders, L. Ioannidou-Mouzaka, et al., Breast cancer screening and diagnosis: a synopsis of the european breast guidelines, Ann. Intern. Med. 172 (1) (2020) 46–56.

[3] M.C. Posso, T. Puig, M.J. Quintana, J. Sola-' Roca, X. Bonfill, Double versus single reading of mammograms in a breast cancer screening programme: a cost-consequence analysis, Eur. Radiol. 26 (9) (2016) 3262–3271.

[4] E.D. Pisano, F. Shtern, Image processing and computer aided diagnosis in digital mammography: a clinical perspective, Int. J. Pattern Recognit. Artif. Intell. 7 (06) (1993) 1493–1503.

[5] E.L. Henriksen, J.F. Carlsen, I.M. Vejborg, M.B. Nielsen, C.A. Lauridsen, The efficacy of using computer-aided detection (cad) for detection of breast cancer in mammography screening: a systematic review, Acta Radiol. 60 (1) (2019) 13–18.

[6] A. Rodriguez-Ruiz, J.-J. Mordang, N. Karssemeijer, I. Sechopoulos, R.M. Mann, Can radiologists improve their breast cancer detection in mammography when using a deep learning based computer system as decision support?, in: 14th International Workshop on Breast Imaging (IWBI 2018), vol. 10718. SPIE, 2018, pp. 7–16.

[7] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, Zorin, S. Jastrzebski, T. Fevry, J. Katsnelson, E. Kim, et al., Deep neural networks improve radiologists' performance in breast cancer screening, IEEE Trans. Med. Imaging 39 (4) (2019) 1184–1194.

[8] K. Lång, V. Josefsson, A.-M. Larsson, S. Larsson, C. Hogberg, H. Sartor, S. Hofvind, I. Andersson, A. Rosso, 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. 24 (8) (2023) 936–944.

[9] O. Diaz, A. Rodrıguez-Ruiz, A. Gubern-Merida, R. Martı, M. Chevalier, Are artificial intelligence systems useful in breast cancer screening programmes? Radiologıa (English Edition) 63 (3) (2021) 236–244.

[10] X. Tang, The role of artificial intelligence in medical imaging research, BJR— Open 2 (1) (2019) 20190031.

[11] J. Li, J.-h. Cheng, J.-y. Shi, F. Huang, Brief introduction of back propagation (bp) neural network algorithm and its improvement, in: Advances in computer science and information engineering, Springer, 2012, pp. 553–558.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017).

[13] C. Matsoukas, J.F. Haslum, M. Soderberg, K. Smith, Is it time to replace cnns with transformers for medical images? arXiv preprint arXiv:2108.09038, 2021.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Proces. Syst. 25 (2012).

[15] H. Greenspan, B. Van Ginneken, R.M. Summers, Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique, IEEE Trans. Med. Imaging 35 (5) (2016) 1153–1159.

[16] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sanchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[17] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221.

[18] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, G. Broeders, P. Gennaro, T. H. Clauser, M. Helbich, T. Chevalier, T.M. Tan, et al., Standalone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists, JNCI: J. Natl. Cancer Inst. 111 (9) (2019) 916–922.

[19] K. Lång, Cancer detection in relation to type and stage in the randomised Mammography Screening with Artificial Intelligence trial (MASAI), European Congress of Radiology, Vienna, Austria, 2024.

[20] R. Agarwal, O. Dıaz, M.H. Yap, X. Llado, R. Marti, Deep learning for mass detection in full field digital mammograms, Comput. Biol. Med. 121 (2020) 103774.

[21] R. Agarwal, O. Diaz, X. Llado, M.H. Yap, R. Martı, Automatic mass detection in mammograms using deep convolutional neural networks, J. Med. Imaging 6 (3) (2019) 031409.

[22] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography, Sci. Rep. 9 (1) (2019) 1–12.

[23] E.F. Conant, A.Y. Toledano, S. Periaswamy, S.V. Fotin, J. Go, J.E. Boatsman, J. W. Hoffmeister, Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis, Radiol. Artif. Intell. 1 (4) (2019) pp.

[24] R.K. Samala, H.-P. Chan, L. Hadjiiski, M.A. Helvie, J. Wei, K. Cha, Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography, Med. Phys. 43 (12) (2016) 6654–6666.

[25] P. Herent, B. Schmauch, P. Jehanno, O. Dehaene, C. Saillard, C. Balleyguier, J. Arfi-Rouche, S. Jegou, Detection and characterization of mri breast lesions using deep learning, Diagn. Interv. Imaging 100 (4) (2019) 219–225.

[26] M.A. Waller, S.E. Fawcett, Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management, 2013, pp. 77–84.

[27] R. Kitchin, The data revolution: Big data, open data, data infrastructures and their consequences, Sage (2014).

[28] CrowdFlower, "Data Science," Tech. Rep., 2016. [Online]. Available: https://visit.figure-eight.com/rs/416-ZBE-142/ images/CrowdFlower DataScienceReport 2016.pdf.

[29] O. Diaz, K. Kushibar, R. Osuala, A. Linardos, L. Garrucho, L. Igual, P. Radeva, F. Prior, P. Gkontra, K. Lekadir, Data preparation for artificial intelligence in medical imaging: a comprehensive guide to open-access platforms and tools, Physica medica, 83 (2021) 25–37.

[30] B. Alyafi, O. Diaz, P. Elangovan, J.C. Vilanova, J. del Riego, R. Marti, Quality analysis of dcgan-generated mammography lesions, in: 15th International workshop on breast imaging (IWBI2020), vol. 11513. SPIE, 2020, pp. 80–85.

[31] B. Alyafi, O. Diaz, R. Marti, Dcgans for realistic breast mass augmentation in x-ray mammography, in: Medical Imaging 2020: Computer-Aided Diagnosis, vol. 11314. SPIE, 2020, pp. 473–480.

[32] R. Osuala, K. Kushibar, L. Garrucho, A. Linardos, Z. Szafranowska, S. Klein, B. Glocker, O. Diaz, K. Lekadir, Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging, Med. Image Anal. (2022) 102704.

[33] Z. Szafranowska, R. Osuala, B. Breier, K. Kushibar, K. Lekadir, O. Diaz, Sharing generative models instead of private data: a simulation study on mammography patch classification, in: 16th International Workshop on Breast Imaging (IWBI2022), vol. 12286. SPIE, 2022, pp. 169–177.

[34] R. Osuala, G. Skorupko, N. Lazrak, L. Garrucho, E. Garcıa, S. Joshi, S. Jouide, M. Rutherford, F. Prior, K. Kushibar, et al., medigan: a python library of pretrained generative models for medical image synthesis, J. Med. Imaging 10 (6) (2023) 061403.

[35] L. Garrucho, K. Kushibar, R. Osuala, O. Diaz, A. Catanese, J. del Riego, M. Bobowicz, F. Strand, L. Igual, K. Lekadir, High-resolution synthesis of high-density breast mammograms: application to improved fairness in deep learning based mass detection, Front. Oncol. 12 (2023) 1044496.

[36] D. Pinto Dos Santos, D. Giese, S. Brodehl, S. Chon, W. Staab, R. Kleinert, D. Maintz, B. Baeßler, Medical students' attitude towards artificial intelligence: a multicentre survey, Eur. Radiol. 29 (4) (2019) 1640–1646.

[37] O. Diaz, G. Guidi, O. Ivashchenko, N. Colgan, F. Zanca, Artificial intelligence in the medical physics community: an international survey, Phys. Med. 81 (2021) 141–146.

[38] J. van Hoek, A. Huber, A. Leichtle, K. Härmä, Hilt, H. von Tengg-Kobligk, J. Heverhagen, A. Poellinger, A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over, Eur. J. Radiol. 121 (2019) 108742.

[39] G. Currie, A muggles guide to deep learning wizardry, Radiography (2021).

[40] K.G. van Leeuwen, S. Schalekamp, M.J. Rutten, B. van Ginneken, M. de Rooij, Artificial intelligence in radiology: 100 commercially available products and their scientific evidence, Eur. Radiol. 31 (6) (2021) 3797–3804.

[41] A. HLEG, Assessment list for trustworthy artificial intelligence (altai) for self-assessment, High Level Expert Group on Artificial Intelligence. B-1049 Brussels, 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/altai-self-assessment.

[42] G.S. Collins, P. Dhiman, C.L.A. Navarro, J. Ma, L. Hooft, J.B. Reitsma, P. Logullo, A. L. Beam, L. Peng, B. Van Calster, et al., Protocol for development of a reporting guideline (tripod-ai) and risk of bias tool (probast-ai) for diagnostic and prognostic prediction model studies based on artificial intelligence, BMJ Open 11 (7) (2021) e048008.

[43] J. Mongan, L. Moy, C.E. Kahn Jr, Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers, Radiol. Artif. Intell. 2 (2) (2020) pp.

[44] X. Liu, S.C. Rivera, D. Moher, M.J. Calvert, A.K. Denniston, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension, BMJ 370 (2020).

[45] L. Maier-Hein, B. Menze et al., Metrics reloaded: Pitfalls and recommendations for image analysis validation, arXiv. org, no. 2206.01653, 2022.

[46] K. Lekadir, A. Feragen, A.J. Fofanah, A.F. Frangi, A. Buyx, A. Emelie, A. Lara, A.R. Porras, A. Chan, A. Navarro, et al., FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare, arXiv preprint arXiv: 2309.12325, 2023.

[47] L. Garrucho, K. Kushibar, S. Jouide, O. Diaz, L. Igual, K. Lekadir, Domain generalization in deep learning based mass detection in mammography: a large-scale multi-center study, Artif. Intell. Med. 132 (2022) 102386.

[48] M.J. Sheller, B. Edwards, G.A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R.R. Colen, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (1) (2020) 1–12.

[49] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Seroussi, Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach, Artif. Intell. Med. 94 (2019) 42–53.

[50] M.A. Mazurowski, "Do we expect more from radiology ai than from radiologists?" Radiology, Artif. Intell. 3 (4) (2021) pp.

[51] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, arXiv preprint arXiv:2006.00093, 2020.

[52] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[53] S. Pacile, J. Lopez, P. Chone, T. Bertinotti, J.M.' Grouin, P. Fillard, Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool, Radiol. Artif. Intell. 2 (6) (2020) e190208.

[54] A. Rodrıguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S.H. Heywang-Kobrunner, I. Sechopoulos, R.M. Mann, Detection of breast cancer with mammography: effect of an artificial intelligence support system, Radiology 290 (2) (2019) 305–314.
[55] S.L. van Winkel, A. Rodrıguez-Ruiz, L. Appelman, A. Gubern-Merida, N. Karssemeijer, J. Teuwen, A.J. Wanders, I. Sechopoulos, R.M. Mann, Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study, Eur. Radiol., 31(11) (2021) 8682–8691.
[56] M.C. Pinto, A. Rodriguez-Ruiz, K. Pedersen, S. Hofvind, J. Wicklein, S. Kappler, R. M. Mann, I. Sechopoulos, Impact of artificial intelligence decision support using deep learning on breast cancer screening interpretation with single-view wideangle digital breast tomosynthesis, Radiology 300 (3) (2021) 529–536.
[57] H.-E. Kim, H.H. Kim, B.-K. Han, K.H. Kim, K. Han, H. Nam, E.H. Lee, E.-K. Kim, Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study, The Lancet Digital Health 2 (3) (2020) e138–e148.
[58] K. Dembrower, A. Crippa, E. Colon, M. Eklund, F. Strand, Artificial intelligence for breast cancer detection in screening mammography in sweden: a prospective, population-based, paired-reader, noninferiority study, The Lancet Digital Health (2023).
[59] M. Larsen, C.F. Aglen, S.R. Hoff, H. LundHanssen, S. Hofvind, Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122,969 screening examinations, Eur. Radiol. 32 (12) (2022) 8238–8246.
[60] N. Sharma, A.Y. Ng, J.J. James, G. Khara, Ambrozay, C.C. Austin, G. Forrai, G. Fox, B. Glocker, A. Heindl, et al., Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms, BMC Cancer 23 (1) (2023) 1–13.
[61] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu, Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis, The Lancet Digital Health 4 (7) (2022) e507–e519.
[62] K. Lång, M. Dustler, V. Dahlblom, A. Åkesson, I. Andersson, S. Zackrisson, Identifying normal mammograms in a large screening population using artificial intelligence, Eur. Radiol. 31 (2021) 1687–1692.
[63] J.L. Raya-Povedano, S. Romero-Martın, E. ElıasCabot, A. Gubern-Merida, A. Rodrıguez-Ruiz, M. Alvarez-Benito, Ai-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation, Radiology 300 (1) (2021) 57–65.
[64] A.D. Lauritzen, A. Rodrıguez-Ruiz, M.C. von Euler-Chelpin, E. Lynge, I. Vejborg, M. Nielsen, N. Karssemeijer, M. Lillholm, An artificial intelligence–based mammography screening protocol for breast cancer: outcome and radiologist workload, Radiology 304 (1) (2022) 41–49.
[65] K. Dembrower, E. Wåhlin, Y. Liu, M. Salim, K. Smith, P. Lindholm, M. Eklund, F. Strand, Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study, The Lancet Digital Health 2 (9) (2020) e468–e474.
[66] Y. Shoshan, R. Bakalo, F. Gilboa-Solomon, V. Ratner, E. Barkan, M. Ozery-Flato, M. Amit, D. Khapun, E.B. Ambinder, E.T. Oluyemi, et al., Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis, Radiology 303 (1) (2022) 69–77.
[67] M. Salim, E. Wåhlin, K. Dembrower, E. Azavedo, T. Foukakis, Y. Liu, K. Smith, M. Eklund, Strand, External evaluation of 3 commercial arti-ficial intelligence algorithms for independent assessment of screening mammograms, JAMA Oncol. 6 (10) (2020) 1581–1588.
[68] Breast Cancer Surveillance Consortium, NCIfunded Breast Cancer Surveillance Consortium co-operative agreement, 2012, www.bcscresearch.org/.