Seminars in Cancer Biology 72 (2021) 214–225

Contents lists available at ScienceDirect

Seminars in Cancer Biology

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

Review

Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art

Ioannis Sechopoulos [a,b,*], Jonas Teuwen [a,c], Ritse Mann [a,d]

[a] Department of Medical Imaging, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA, Nijmegen, the Netherlands
[b] Dutch Expert Centre for Screening (LRCB), Wijchenseweg 101, 6538 SW, Nijmegen, the Netherlands
[c] Department of Radiation Oncology, Netherlands Cancer Institute (NKI), Plesmanlaan 121, 1066 CX, Amsterdam, the Netherlands
[d] Department of Radiology, Netherlands Cancer Institute (NKI), Plesmanlaan 121, 1066 CX, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Artificial intelligence
Screening
Mammography
Tomosynthesis
Breast cancer

ABSTRACT

Screening for breast cancer with mammography has been introduced in various countries over the last 30 years, initially using analog screen-film-based systems and, over the last 20 years, transitioning to the use of fully digital systems. With the introduction of digitization, the computer interpretation of images has been a subject of intense interest, resulting in the introduction of computer-aided detection (CADe) and diagnosis (CADx) algorithms in the early 2000's. Although they were introduced with high expectations, the potential improvement in the clinical realm failed to materialize, mostly due to the high number of false positive marks per analyzed image.

In the last five years, the artificial intelligence (AI) revolution in computing, driven mostly by deep learning and convolutional neural networks, has also pervaded the field of automated breast cancer detection in digital mammography and digital breast tomosynthesis. Research in this area first involved comparison of its capabilities to that of conventional CADe/CADx methods, which quickly demonstrated the potential of this new technology. In the last couple of years, more mature and some commercial products have been developed, and studies of their performance compared to that of experienced breast radiologists are showing that these algorithms are on par with human-performance levels in retrospective data sets. Although additional studies, especially prospective evaluations performed in the real screening environment, are needed, it is becoming clear that AI will have an important role in the future breast cancer screening realm. Exactly how this new player will shape this field remains to be determined, but recent studies are already evaluating different options for implementation of this technology.

The aim of this review is to provide an overview of the basic concepts and developments in the field AI for breast cancer detection in digital mammography and digital breast tomosynthesis. The pitfalls of conventional methods, and how these are, for the most part, avoided by this new technology, will be discussed. Importantly, studies that have evaluated the current capabilities of AI and proposals for how these capabilities should be leveraged in the clinical realm will be reviewed, while the questions that need to be answered before this vision becomes a reality are posed.

1. Breast cancer screening and diagnosis

Every year, over half a million women die of breast cancer worldwide [1]. To reduce the breast cancer-related mortality, screening for breast cancer with mammography has been introduced in many countries around the world over the last three decades. Screening, together with improvements in treatment, has resulted in a reduction in breast cancer mortality of ~30 % [2], but this disease is still the number one cause of female cancer death [1].

Breast cancer screening with mammography has been implemented differently in different countries. In many countries, like in the US, screening is institution-based. Women, by themselves or referred by

Abbreviations: AI, artificial intelligence; AUC, Area under the receiver operating characteristics curve; CADe, computer-aided detection; CADx, computer-aided diagnosis; CC, cranio-caudal; CNN, convolutional neural network; DBT, digital breast tomosynthesis; DM, digital mammography; LoS, level of suspicion; MLO, mediolateral oblique; PoM, probability of malignancy; ROC, receiver operating characteristics.

* Corresponding author at: P.O. Box 9101, Route 766, 6500 HB Nijmegen, the Netherlands.
E-mail addresses: ioannis.sechopoulos@radboudumc.nl (I. Sechopoulos), jonas.teuwen@radboudumc.nl (J. Teuwen), ritse.mann@radboudumc.nl (R. Mann).

https://doi.org/10.1016/j.semcancer.2020.06.002
Received 7 November 2019; Received in revised form 19 May 2020; Accepted 1 June 2020
Available online 9 June 2020

their primary care physician or gynecologist, present at a breast imaging center, many times situated in or affiliated with a hospital, for their screening exam. Although heavily regulated, the details of the screening process (digital mammography (DM) and/or digital breast tomosynthesis (DBT), manufacturer, use of computer-aided interpretation, etc.) is decided by the institution. Depending on the institution, the acquired images may be interpreted while the woman waits, and any additional imaging is performed during the same visit. At larger screening centers, the screening exams are read in batches, one or two days after acquisition, and if a suspicious finding is detected, the woman needs to be *recalled*. In many countries, especially in Europe, breast cancer screening has been implemented as a government (regional or national) program. In these programs, women of a certain age range (commonly 50–70 years old) receive an invitation to get their mammographic screening exam periodically (commonly every two years). Some screening programs have their own dedicated screening centers, un-affiliated with any hospital. In general, there is a larger (or complete) degree of homogeneity in the equipment and processes used in these programs. In screening programs, the exams are batch-read, and recalls are actually denoted *referrals*, since usually the case is forwarded to a hospital, for further imaging and testing.

When radiologists are interpreting screening mammograms, they are searching for lesions with very different characteristics that can be divided into two broad categories: calcification clusters and soft tissue findings. The calcifications of interest for the detection of breast cancer are small (as little as 0.2 mm) and relatively high in contrast. The shape of the calcifications and the distribution of the cluster of calcifications being important biomarkers for malignancy. Soft tissue lesions are of different types; masses (with different shape and margin descriptors, such as spiculated, smooth, obscured, irregular), architectural distortions (abnormal configuration of the fibroglandular tissue) and asymmetries (dense tissue patterns in one breast with no correspondence on the contralateral breast).

Of course, in the detection of breast cancer, one major biomarker for the presence of malignancy is a change (for the most part, growth) in the finding itself. In other words, a suspicious finding that is found to not change with time is usually deemed as not of concern. Therefore, during interpretation of screening mammograms, the comparison to the prior images is important, in improving both sensitivity and specificity [3–6], and provides additional information different from that gained by other concurrent imaging, such as digital breast tomosynthesis (DBT, described later) [7].

In many screening programs, especially in Europe, each case is reviewed by two radiologists (usually independently), a process called *double reading*. Each reader interprets the images and decides whether the woman needs to be recalled/referred for further evaluation of a suspicious finding. If the two readers do not agree in their assessment, depending on the program setup, either they meet to arrive at a *consensus*, or a third radiologist acts as an *arbiter,* whose opinion prevails. Although double reading requires more resources than single reading (the common process in the US), it has been shown to improve the cancer detection rate at screening, although it also increases the recall rate, resulting in a comparable positive predictive value [8,9].

Aside from double vs. single reading, another major difference between screening in Europe and in the US (and some other countries), is the recall rate, i.e. the proportion of screened women that are recalled or referred for further testing. As extreme examples, the referral rate in the Netherlands and Sweden is about 2.5 % [10,11], while in the US the recall rate is about 11.5 % [12] (now ~30 % lower with DBT). This difference may be attributed, to differences in practice and to the medico-legal implications of a missed cancer.

After screening, when a woman is recalled, or referred, she undergoes *diagnostic work-up*, to determine if the suspicious finding at screening is indeed a lesion of concern. This work-up can consist of additional DM and/or DBT imaging, ultrasound, and, in some limited cases, contrast-enhanced breast MRI. Based on this additional imaging, the interpreting radiologist decides if a biopsy is warranted, or if the finding was a false positive. If a biopsy is performed, depending on the nature of the lesion, this could be done using fine-needle aspiration, core or vacuum-assisted biopsy, or, in rare cases, an excisional biopsy. Final diagnosis is done based on the pathological analysis of the biopsy-obtained sample, which, in these cases, determines if the screening assessment was a true or false positive. If the screening was assessed as normal, guidelines state that this assessment is considered a true or false negative depending on the woman having had breast cancer diagnosed or not during the period in between screening rounds.

2. Digital mammography

In its first implementation, breast cancer screening was performed with screen-film mammography. Since the early 2000′s, with the introduction of affordable large-area digital detectors, digital mammography (DM) was developed and introduced for clinical use. In DM, the use of film was replaced with a digital x-ray detector, which would immediately result in a digital image, ready for evaluation for appropriateness by the acquiring radiographer, and interpretation by the radiologist. An intermediate, alternative pathway to digitization of the breast screening process is the use of computed radiography-based mammography. However, various studies have shown the inferior performance of this technology, and therefore its use is being reduced [13–15].

DM has various advantages over screen-film mammography, chief among them the simpler workflow. In terms of performance, DM has been shown to have improved clinical performance in sub-groups of the screening population [14,16], but also being equivalent to screen-film mammography in the general screening population [17–20]. Beyond these improvements, one additional advantage of the introduction of digital detectors for breast imaging is the ease with which the technology can be extended by developing more advanced image acquisition methods, such as DBT and dedicated breast CT, as well as the introduction of post-acquisition processing and analysis algorithms.

Mammography, both screen-film and digital, involves the acquisition of a single two-dimensional image of the breast. This results in the phenomenon of tissue superposition, in which different tissues in the breast, separated only in the direction of the projection, are projected onto the same location in the 2D mammographic image (Fig. 1). As a result, normal tissues may cover up the presence of a malignant lesion, reducing sensitivity, and, the projection of separate normal tissues may mimic a suspicious lesion, reducing specificity. These effects substantially reduce the accuracy of 2D mammography, especially in breasts with a large amount of fibroglandular tissue (i.e. dense breasts), which is present in about half of the screened breasts [21], and is responsible for one third of missed cancers [22].

To alleviate the issue of tissue superposition and loss of performance in dense breasts, screening mammography is performed by acquiring two views of each breast: the cranio-caudal (CC) and the medio-lateral oblique (MLO) views. These two views are evaluated together by the interpreting radiologist, in a cognitive effort to determine if a candidate lesion seen in one view is present in the other, or can be discarded as random tissue superposition, in addition to the hope that a different breast compression direction results in an otherwise occult lesion being seen in at least one of the views.

In summary, interpretation of screening mammograms includes the review and comparison of features (or lack thereof) across views of the same breast and across images of the two breasts at the current timepoint, as well as comparison of images acquired at separate timepoints. Ideally, to maximize performance, any automated image evaluation algorithm for detecting breast cancer at screening should be capable of performing these same comparisons.

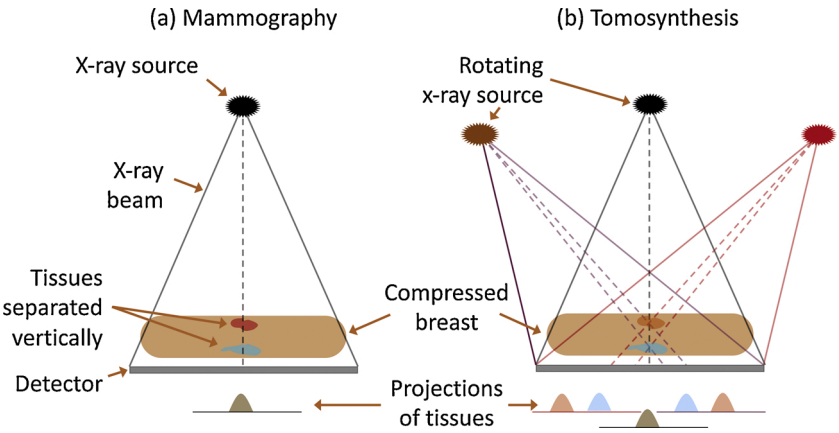

Fig. 1. Diagram of a (a) digital mammogram and a (b) digital breast tomosynthesis acquisition. Tissues in the breast that are only separated in the vertical direction appear superimposed in the mammogram, resulting in a loss of sensitivity and specificity. This effect is ameliorated in digital breast tomosynthesis by reconstructing a pseudo-3D image from several projections, each acquired with the x-ray source positioned at a different angle.

3. Digital breast tomosynthesis

Owing to the limitations in DM due to its two-dimensional nature, the last two decades has witnessed the development, and over the last decade, the clinical introduction, of digital breast tomosynthesis [23] (Fig. 1). DBT is a pseudo-tomographic imaging technique that results in a stack of 2D slices of the imaged breast, with some, albeit limited, vertical resolution. This partial tomographic effect reduces the masking effect of superimposed tissues. Studies have reported an increase in cancer detection with a, mostly, lowering of the recall rate, depending on what the baseline (DM) recall rate was [11,24–29].

Although trials have resulted in promising results in terms of cancer detection with DBT at screening, one major drawback is its increase in interpretation time compared to DM. It has been consistently reported that, due to the substantial increase in the number of images needed to be reviewed, interpretation of DBT images takes approximately double the time required for reading DM images [28]. As a result, introduction of DBT in large-scale screening programs will be dependent on not only its impact on clinical outcomes, but also in the introduction of methods to reduce its reading time. Automated methods of interpreting these images will surely have an impact in the potential for introduction of DBT for screening. This impact could be two-fold; in the first place, computer-driven faster navigation of the DBT image stack could result in important reductions in the time spent by the radiologist in the visual search for suspicious findings. In addition, the use of computer methods to aid in interpreting the DBT images could reduce the variability seen across studies in the impact of DBT on recall rate at screening, if inter-reader variance is reduced.

4. Conventional computer-aided detection and diagnosis

The possibility of digitizing screen-film mammograms, and then the introduction of DM into the clinical realm, resulted in an expanding interest in leveraging the use of computers to aid in the interpretation of screening mammograms. Two categories of computer algorithms were investigated and developed; computer-aided detection (CADe) and computer-aided diagnosis (CADx).

CADe is aimed at locating suspicious lesions, either soft tissue masses and/or calcification clusters. All conventional CADe algorithms are based on the same three-part strategy: (i) normalize the image to a "reference" intensity distribution (usually an arbitrary intensity distribution that the following steps have been prepared for) and/or process the image to enhance the detectability of suspicious signals, (ii) identify areas of the image with candidate suspicious signals, and (iii) reduce the number of identified regions by evaluating the probability of an actual lesion being present in each region, and applying a threshold to this probability [30]. CADx algorithms estimate if a given, already-detected lesion is benign or malignant, and therefore involve a similar approach as the final step of a CADe process, albeit without the use of a threshold.

To identify and rate the suspiciousness of lesions, conventional CADe/CADx systems use programmed-in features; the algorithms are programmed to search for specific features that humans have identified as representative of suspicious lesions. As we will see later, this is the primary distinguishing feature between conventional CADe/CADx algorithms and current, state-of-the-art AI-based algorithms.

To improve the performance of CADe algorithms, just as done by radiologists, algorithms have been developed to review the information contained across the two different (CC & MLO) views of the same breast, and across the matching views of the two breasts. Engeland and Karssemeijer developed an algorithm to detect and evaluate lesions across the two views of the same breast, and incorporated it into a previously developed CADe program [31]. Tahmoush and Samet, and Wang et al., proposed algorithms to detect asymmetries across the corresponding views of the two breasts, resulting, as expected, in a substantial improvement in the performance of CADe [32–34].

CADe was introduced for use during screening to reduce the frequency of lesions that are overlooked by the interpreting radiologist, as a *second reader*. That is, after the interpreting radiologist reviews the entire case and arrives at a decision, he or she would turn on the CADe, and determine if any of the computer-generated marks are of concern or not [35]. Upon introduction of these algorithms, great promise for improved outcomes was foreseen, with encouraging initial results [22,36–38]. However, studies that showed improved performance with the use of CADe were mostly either focused on specific types of lesions or evaluated small, enriched, data sets. After years of clinical use, large scale retrospective analysis of the impact of the introduction of CADe on screening performance indicated that the expected benefits of CADe did not materialize [39,40]. Overall, the use of CADe was determined to lower specificity and positive predictive value (the probability of disease present given a positive test), while not resulting in a significant increase in sensitivity. In fact, in a sub-group analysis of radiologists that had access to CADe only a for a portion of their interpretations (due to working at more than one site), sensitivity was lower when CADe was used [40]. This points to the possibility that in many cases the CADe algorithm is not used as a second reader, but rather as a first reader, with the radiologist then only reviewing the marks for deciding to recall or not.

The major pitfall of conventional CAD is the rate of false positive marks. Good CADe sensitivity performance is achieved only when setting the internal CADe threshold for marking suspicious areas at rates

above one finding per image. Considering the actual prevalence of cancer in a screening population, <1% [41,42], or even the recall rate at screening, ~2.4 %–11.5 % [12,42], depending on the country, it is obvious that the great majority of these marks are false positives not only in terms of malignancy, but also as actionable findings. Therefore, even if a few of these marks prompt the interpreting radiologist to initiate a recall, then specificity will decline. As far as reducing the number of actual overlooked cancers, over 1 000 false positive marks need to be considered for one additional cancer to be detected [43].

Conventional CAD performance on DBT should not be expected to be better, given that, for example, a commercial CADe product for DBT was reported to have a per-lesion sensitivity of 89 %, with a 2.7 ± 1.8 false-positive rate per view [25]. Several other reported performances for CAD for DBT images are also in the range of 1 or more false positive marks per view [44].

In summary, conventional CAD, for both DM and DBT, has not reached a level of performance that could improve actual screening performance in the real world, despite early hopes and promises and years of clinical use [39,40,45].

5. Deep learning convolutional neural networks for medical image classification

The introduction of deep learning convolutional neural networks (CNNs) in medical image analysis has brought forth a potential revolution in computer-based interpretation of DM and DBT images. The important developments in the last few years in the field are due to the use of these multi-layered CNNs, but, as in the title of this review, it is common to refer to artificial intelligence (AI) and deep learning almost interchangeably. However, these terms are not synonyms; AI includes many different types of techniques. Within AI lies machine learning, which includes deep learning of which, finally, CNNs are only a subset [46] (Fig. 2).

Deep learning convolutional neural networks involve the processing of an image by multiple, sequential, stages, denoted *layers*, of usually simple multiplication, addition and maximum (convolutions and downsampling) mathematical operators, that combine the spatially correlated information contained in images. During this multiple-stage process, this information is broken down into different representations, and the analysis of these more abstract, and simpler, representations of this information results in the ability of the network to recognize the image accurately.

Deep learning CNNs first made an impact in image classification when the submission by Krizhevsky et al. won the 2012 ImageNet Large Scale Visual Recognition Challenge by a landslide [47]. Since then, interest in this technology for various image classification applications increased quickly, and its use for detection of breast cancer in mammography has been investigated for the last few years. A comprehensive survey of the initial introduction of deep learning in medical image analysis, including in the field of breast imaging, has been published by Litjens et al. [48].

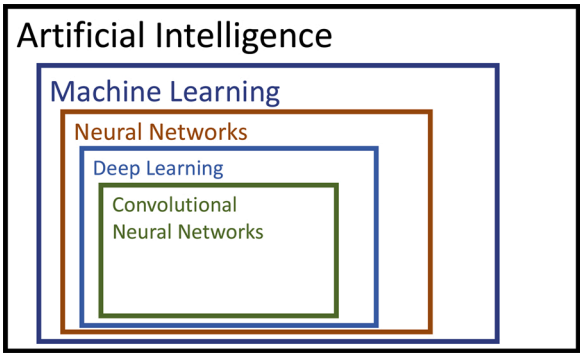

Fig. 2. Diagram explaining the relationship between the different methods and algorithms in the field of artificial intelligence.

The major characteristic that distinguishes this new AI-based image classification algorithms from conventional CAD is that the determination of what image features are indicative of a lesion being present is achieved by the algorithm itself during its training, not input by the human programmer. In other words, the algorithm is not taught what a breast cancer looks like (size, shape, texture patterns, etc.) but it teaches itself what it looks like. This is achieved during the training process, by providing the model many examples of images (portions or complete images) with and without cancers present, each of them labeled with its actual status (cancer present/not present). During the training, for each input example image the deep learning network adjusts its internal parameter values to minimize the difference between its predicted status of the image to that of the truth. In this manner, the network recognizes what the image features are that point to a malignant lesion being present.

One simplifying aspect of obtaining training data for AI algorithms in breast cancer imaging is that the true status of the case is, relatively speaking, straightforward. As opposed to other pathologies, e.g. many cardiac diseases, the determination that a mammogram contains a malignant or benign lesion, or no lesion at all, follows a well-accepted standard, and the vast majority of studies conform with this, perhaps unwritten, rule. For images containing lesions, their malignant or benign status should be confirmed by pathological analysis of biopsy samples, while normal cases normally include one- or two-year follow-up with no cancer diagnosis. In some studies, cases including obviously benign lesions may have not been biopsied, but their benign status was confirmed by long-term follow-up. All of the following studies mentioned here have used this definition of truth.

A distinction should be made between image-level and pixel-level classification. Image-level classification involves identifying an entire image as containing a cancer or not. Pixel-level classification includes determining where in the image the lesion is located, either through providing a region-of-interest or by labelling each pixel in the image as whether belonging to a lesion or not. In screening, the basic task is identifying the person that needs further evaluation due to being at high (-er) risk of having the disease being screened. For example, screening for prostate cancer by measuring the level of prostate-specific antigen in blood, or screening for cervical cancer via a pap test, does not provide additional information on the location or nature of the suspicion. Mammographic screening does result in an indication of where the suspicion is located and its nature (soft tissue mass, calcifications, etc.). Therefore, any automated DM or DBT image evaluation algorithm should provide the location of the detected suspicious finding.

6. AI-based algorithms for breast cancer detection in digital mammography and digital breast tomosynthesis

Because of the special characteristics of screening with DM and DBT, algorithms to automatically detect breast cancer in these images usually go beyond the "standard" deep learning CNNs. In the first place, algorithms for breast cancer detection in DM and DBT need to search for both soft tissue lesions and calcifications. Given their very different characteristics and the frequently still-limited training datasets, usually different separate detection algorithms are used for each of these types of lesions, and the results are combined at the final stage of analysis. For example, Lotter et al. developed a two-stage algorithm, in which first two different multi-scale CNNs, one for masses and the other for calcifications, are used to scan and analyze the image in patches, and then the output of these is aggregated to pool together both across lesion types and analysis scales, resulting in a final classification estimate [49]. However, at least one image-level classification network has been reported on that does not involve separate analysis of the images in search of soft tissue lesions vs. calcifications, resulting in good performance [50].

Also, except for the special application of pre-identification of normal cases (as will be discussed later), the algorithm should identify the location of the suspicious finding(s), not only determine if an image contains a suspicious lesion. This requires algorithms that go beyond standard image classification with DL CNNs. Development of such algorithms has included combining the information gained from analyzing patches using hand-crafted features used in conventional CAD with the deep learning CNN analysis, which has resulted in improved overall performance at the patch level [51]. Instead of combining the conventional feature analysis with the CNN algorithm itself, Samala et al. [52] used it as a pre-screening stage to identify suspicious areas of clustered calcifications, and then designed a deep learning CNN to differentiate the true calcifications from false positives, resulting in improved performance over the use of only a deep learning CNN. Therefore, it seems that although the next-generation deep learning CNNs result in improved performance over conventional CAD, the use of the information gained from the latter, combined with the former, does result in an even higher performing system.

Some work, however, has been performed on using a single CNN, involving the aptly-named algorithm *YOLO* (You Only Look Once) that analyzes the entire image, without the use of more complex, multi-stage or multi-network algorithms, performing both detection and characterization, resulting in identification of not only the presence of lesions but also providing location information [53]. Comparison of its performance to conventional CAD methods shows, not surprisingly, considerably improved outcomes. Furthermore, its performance compared to a patch-based CNN analysis for characterization only on the same data set was shown to be equivalent. How this method compares to the other approaches, including the combined deep learning-conventional feature analysis method described above, remains to be seen.

It should be noted that methods that include analysis at the pixel- or patch-level usually require annotated training sets, in which the malignant lesions are outlined in the images, or the images consist of only the image patches where the lesions are located. This greatly increases the difficulty in obtaining adequate training datasets, since annotation of images is a lengthy, tedious process, that needs to be performed by subject-matter experts, and is still fraught with inter-reader variability [54]. Therefore, minimizing the amount of training that these algorithms require, and therefore the size of the needed sets, is of interest. One effective way to achieve this is *transfer learning* [55]. Transfer learning involves using an already-trained deep learning CNN, keeping a significant portion of the internal CNN parameter values constant, and only fine-tuning the parameters of the final layers of the network for the new application. In this way, the information from unrelated, very large data sets, like the natural image set ImageNet, consisting of over a million images, can be leveraged in the training of a CNN that is oriented at analyzing mammograms [56–58]. For example, Samala et al., starting from the pre-trained deep learning CNN AlexNet, successfully fine-tuned the network for breast cancer detection in both analog and digital mammography with only ~1500 lesion image patches, of which only 500 were analog images and 96 were digital images containing malignancies [56]. Considering that the original training of AlexNet consists of using over 1.2 million natural (non-medical) images, the power of transfer learning is very significant.

In addition to taking advantage of transfer learning, other methods have been proposed to reduce the complexity of training of these algorithms, including decreasing their complexity, and therefore the number of parameters needing to be tuned, by *pruning* of the CNN, in manners that do not affect its performance. In another study, Samala et al. [59] were able to reduce the complexity of a deep learning CNN by approximately 90 % in number of neurons and mathematical operations needed, and by one third in number of parameters needing to be set, resulting in the same DBT lesion detection performance.

Further improvements in AI algorithms for lesion detection have been proposed by developing methods to analyze the DM and DBT images across the two breasts, and, finally, comparison with prior exams. For this, Kim et al. developed a deep learning CNN-based method to analyze bilateral views in DBT to detect masses, resulting in improved performance over the use of hand-crafted features [60]. Kooi and Karssemeijer also developed methods to detect asymmetries by comparing the images acquired across breasts, and methods to compare current exams to prior ones [61,62]. A significant improvement was found in the performance of the detection algorithm by the inclusion of the asymmetry analysis when operating at high specificity (and therefore resulting in a low recall rate in screening), as would be important. However, surprisingly, no significant improvement was found by the incorporation of the temporal comparison, as would be expected given its impact on clinical practice [3–6]. Therefore, either more development needs to be accomplished in this area, or there are present as-yet unknown biomarkers in the current images, equivalent to the information provided in the prior images, that deep learning algorithms have been able to identify, rendering the addition of the prior information of little or no value.

In what could result in a very interesting step forward for the diagnosis of breast cancer, Hamidinekoo et al. proposed a new framework consisting of linking a pair of deep learning models, one that analyzes DM images and the other for analysis of histopathology slides, each independently trained by the corresponding human-annotated ground truth [63]. Each model would be used to identify phenotypes present in its own data source. The outcomes of both analyzes could then be matched, resulting in an objective, computer-based model of the correlation and association between the DM imaging data and the pathology data.

7. AI for digital mammography vs. AI for digital breast tomosynthesis

Given the similarity between the two imaging modalities, extensive work has been performed to understand how the AI developments for analysis of DM images relate to those in DBT, and if and how some of the knowledge, training data, and methods developed for one can be applied to the other. For example, Zhang et al. tested various standard CNNs for image-level classification on both DM and DBT images, using similar methods, showing that transfer learning and other advanced training methods were feasible to apply to develop systems for both modalities [64].

In a similar way as networks trained using natural images can be fine-tuned, via transfer learning, for use on DM images, it has been found that networks trained for DM can be fine-tuned for DBT. Specifically, a network was initially trained on DM images of 2461 lesions (from which 45,072 patches were generated after data augmentation), that could then be fine-tuned, via transfer learning, using only 228 lesions on DBT (resulting in 37,450 patches) [65]. The performance of this fine-tuned network was even better than that obtained by DBT-specific conventional CAD methods. This is an important determination, given the much more extensive databases and clinical archives that exist of DM images compared to those available of DBT exams. It should be noted that previously, work had been performed to show that conventional CAD algorithms for DM can be used successfully on reconstructed DBT slabs [66]. Therefore, that this would also be true for the new, AI-based, detection algorithms, is not unexpected.

Transfer learning from natural-image-trained networks has also been used directly for fine-tuning for DBT and synthetic DM networks, showing good performance for characterization of lesions, i.e. CADx applications [67]. When comparing the performance across all three types of images (DM, center-of-lesion slice of DBT, and synthetic 2D), with an algorithm that combines the information from both views (CC & MLO), the best performing image type was the DBT slice, which is expected since it does not suffer from superposition of fibroglandular tissue.

8. How did these new AI algorithms evolve?

In the days of initial work performed with deep learning CNNs for breast cancer detection in mammography, studies were performed to compare the difference in performance between the current technology, i.e. conventional CADe/CADx, and the up-and-coming CNN-based technology. For example, Fotin et al. compared conventional CADe to the performance of deep learning CNNs for the task of mass detection in DBT [68]. Moving from conventional to deep learning approach resulted in an increase of sensitivity, at the ROI level from 83.2%–89.3% for ROIs marked as containing suspicious lesions, and 85.2%–93.0% for ROIs containing malignant lesions.

In another early study, but with comparison of performance against radiologists instead of against conventional CADe, Becker et al. used a deep learning-based commercial image analysis algorithm intended for industrial use, which is not approved for medical use [69]. The algorithm was trained and tested with two different data sets: one clinical set with a 50 %/50 % proportion of malignant/control cases, and another set with an approximately 10 %/90 % proportion of cases, to better resemble screening, although the prevalence of cancer is still ~10 times higher than in a real screening set. It was shown that deep learning algorithms, even when designed for non-medical imaging purposes, can be trained for the application of detecting breast cancer in DM. For the high prevalence set, two of the three readers performed significantly better than the algorithm, while for the lower prevalence set the algorithm performed comparably to the radiologists.

As described above, Kooi et al. [51] developed an AI system for DM analysis that uses a deep learning CNN in combination with hand-crafted features. In that study, they compared the performance of the new system to that of both a conventional CADe algorithm and to the performance of humans interpreting the same DM images. As can be seen in Fig. 3, the proposed CNN, when having access to only the image patch, with no external information, making the conditions equivalent to those available to the conventional CADe, resulted in a non-significant increase in the area under the receiver operating characteristics (ROC) curve (AUC) compared to latter. However, the performance of the CNN increased with the incorporation of the handcrafted features.

As has been discussed, the major pitfall of conventional CADe in real world use was the number of false positive marks per image. Therefore, beyond the comparison of the overall performance, the new AI-based technology would be expected to make a significant impact on healthcare only if it outperforms conventional CAD in the high specificity region of the ROC curve. In other words, if the impact on the sensitivity of operating at very low false positive marks per image were small. As can be seen in Fig. 4, the decrease in sensitivity with decreasing false positive marks per image is considerably lower for the CNN-based algorithm than for the reference CADe system, resulting, e.g., in a 20 % difference in case-based sensitivity at 0.1 false positive marks per image (1 false positive every 10 images).

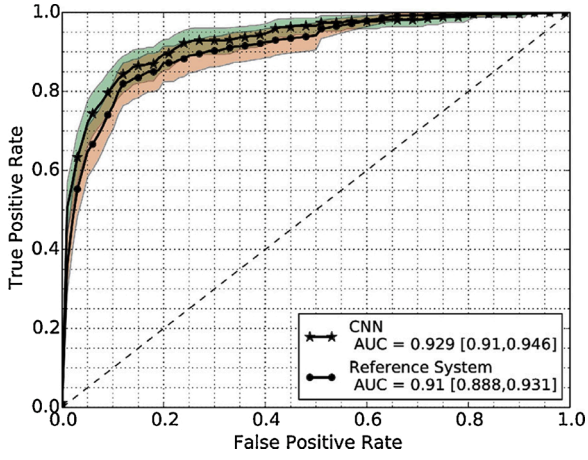

Fig. 3. Stand-alone ROC performance of a deep learning CNN compared to a conventional (reference) CADe, when restricting the CNN to analyze only image patches, with no additional information, to be on par with the information available to the conventional CADe algorithm. Reprinted with permission from Kooi et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical Image Analysis. 2017;35:303–312.

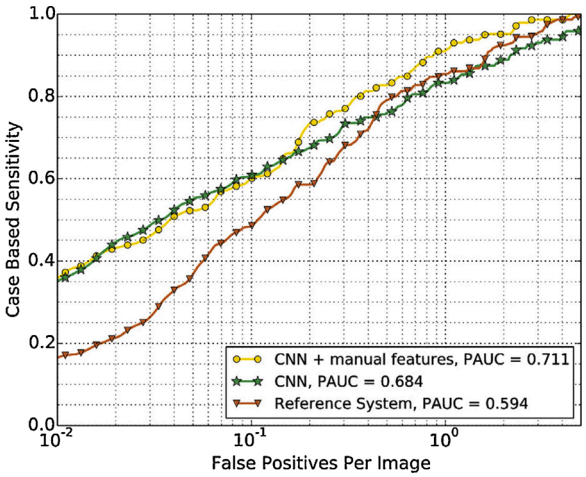

Fig. 4. FROC curve showing the reduction in sensitivity with decreased false positives per image for a deep learning CNN, the deep learning CNN combined with hand-crafted features, and the conventional (reference) CADe algorithm. Reprinted with permission from Kooi et al. Large scale deep learning for computer aided detection of mammographic lesions. Medical Image Analysis. 2017;35:303–312.

In the comparison of the CNN system to humans (including one imaging scientist expert in breast imaging and two breast radiologists), no significant difference was found between the algorithm and each individual human observer. The pooled results from the observers did result in a significant better performance than that of the algorithm. It should be noted that this comparison was done at the image patch level, to avoid the observers from gathering information that the CNN algorithm was not able to take advantage of (e.g. evaluation of lesion presence across views, asymmetries across breasts, etc.), but obviously potentially also limiting maximum performance. As demonstrated by this early study, deep learning CNN-based algorithms had the potential to have an impact in the evaluation of DM images, with the initial studies already showing performance comparable to human performance. Of course, excluding the human observer's ability to leverage the information from all available sources, such as multiple views, bilateral exams, and prior images, made the comparison of performance only indicative of potential and useful to identify limitations that needed to be worked on, rather than a true comparison to determine if CNN algorithms were ready for clinical use.

9. Stand-alone performance of current state-of-art AI algorithms for digital mammography and digital breast tomosynthesis

Currently, the top AI algorithms for lesion detection and classification in DBT and mammography are all based on CNNs. Several companies offer commercial AI applications that have already been approved by the FDA, are CE-marked or are in the final phases of evaluation. To assess the performance of a commercial AI system for DM, Rodriguez Ruiz et al. collected nine data sets from sites in both the US and Europe [70]. These enriched data sets consisted of DM images and the probability-of-malignancy (PoM) ratings for each case given by a number of breast screening radiologists during retrospective ROC observer studies in which DM was being compared to some other imaging modality. As a result of the collection from multiple sites, the final

data set comprised a total of 2 652 exams, of which 653 were malignant, spanning images acquired with systems from four different vendors, interpreted by 101 radiologists, yielding a total of 28 296 independent interpretations by radiologists from both continents. The case-level performance of the AI system was statistically non-inferior to that of the average of the 101 radiologists (Fig. 5). When comparing the performance of the AI system to each individual radiologist, the investigators found that the former performed better than 61 % of the radiologists. As can be seen in Fig. 6, where the ROC curves comparing the individual radiologist performance to that of the AI system for the largest datasets that included and excluded the evaluation of priors during the radiologist reading (the AI is unable to read the prior exams), the ROC performance of the AI system is consistently similar to that of the radiologists, regardless of the operating point.

The major strength of this large-scale study to determine the potential performance of current AI technology for detection of breast cancer in mammography, aside from the high number of cases and readers involved, is the variability in the data. Due to its multi-site nature, the investigators were able to perform this comparison with images acquired with systems from 4 different vendors interpreted by radiologists from 7 different countries. Regarding the latter, it is particularly interesting that radiologists from both US and Europe were included, given the differences that exist in the approach to screening between them. Of course, this study did not answer all questions and is not fully predictive of how a current AI system compares to humans performing actual breast cancer screening interpretation. In the first place, the data sets used were highly enriched and were interpreted in a laboratory setting. In addition, some data sets were unilateral and some excluded priors. As discussed earlier, radiologists rely heavily on the comparison against the contralateral and the prior breast image during their exam interpretation, while current AI systems are not capable of comparing images across time. However, as can be seen in the two graphs in Fig. 6, the performance of the radiologists, relative to that of the AI, when the former had priors available was not affected, as would be expected. It seems that the various factors that condition this study: non-screening disease prevalence, lab setting reading, availability or not of priors, may have had competing influences, and therefore in which direction the results are biased, if any, is hard to establish [71]. Therefore,

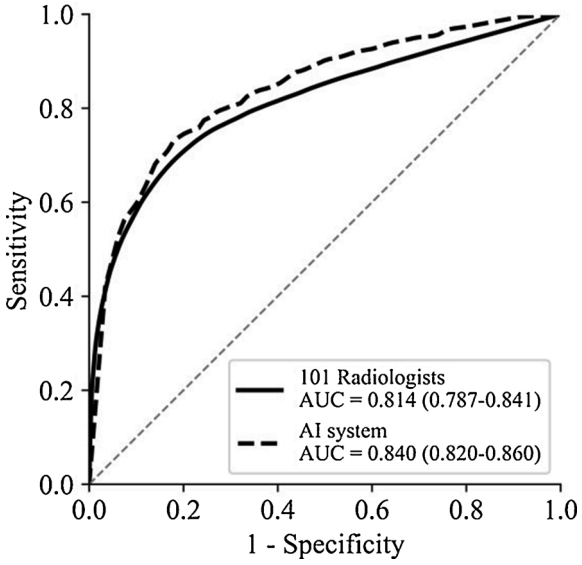

Fig. 5. Stand-alone ROC performance of a commercial AI system compared to the reader-averaged radiologist interpreting over 2 600 mammograms during retrospective observer studies. Reprinted with permission from Rodriguez Ruiz et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916–922.

although this comprehensive study provided very important information regarding the state of the art in AI evaluation of DM, further investigation is needed, ideally including results from large-scale data sets involving radiologist performance at screening.

As mentioned above, whole-image classification methods that are not trained with annotated images require very large training data sets. In addition, these methods require special additional steps to retrieve, for highlighting to the user, the location of the detected suspicious finding. In an example of this type of algorithm, Kim et al. used a data set of over 4 000 cancer cases and almost 25 000 normal cases, all without pixel-level annotations, to train, validate, and test a deep learning CNN that could classify the images into malignant or not, and generate heat maps highlighting the area that most strongly contributed to the final classification decision [50] (Fig. 7). The data set used included screening and diagnostic work-up images from three DM system vendors (although one of them was represented by an order of magnitude fewer cases).

Over the entire test set, the algorithm resulted in an AUC of 0.906, with a sensitivity of 76.1 % at a specificity of 88.5 %. Interestingly, the performance of the algorithm on the images from the manufacturer with the fewest cases was not lower than that for the cases from the other two manufacturers. Although the authors did not include any comparison to human performance, and therefore the actual capability of the system is challenging to judge, the size of the test set and the performance obtained point to the promising potential of this algorithm.

In a study evaluating the use of a commercial AI product by 24 radiologists retrospectively reading an enriched data set of 260 DBT cases, Conant et al. obtained the ROC graph in Fig. 8 when comparing the stand-alone performance of the AI system to that of the radiologists reading without the AI system [72]. The average sensitivity and specificity of the readers were 77.0 % (range: 38.5 %–93.8 %) and 62.7 % (range: 22.1 %–84.6 %), while the corresponding metrics for the AI system were 91 % and 41 % (the operating point of the AI depicted in the ROC graph in Fig. 8). In this study, all DBT cases were acquired with the same system from a single vendor, at seven different sites in the US, and included the accompanying either DM image or synthetic 2D image and excluded the use of priors. All readers practiced in the US and encompassed both general radiologists and breast radiologists. Again, this study was also based on radiologists reading the cases retrospectively, as part of an enriched data set, without the use of priors.

Schaffter et al. reported the results of a grand challenge in which, during the first phase, 31 different AI DM detection algorithms were evaluated using two datasets, one consisting of over 40,000 screening DMs from the US and over 166,000 screening DMs from Sweden [73]. None of the individual algorithms, nor the combination of the 8 best performing algorithms, matched the performance of the original interpreting radiologists. Therefore, although other studies have shown algorithm performance that does match, or surpass, human performance, the size and various phases of this challenge make it of considerable interest. In a subsequent phase, once the combined algorithms were also merged with the radiologist decision, was a significant improvement possible compared to the radiologists alone. Interestingly, again in this report there was no real benefit seen when additional information, such as clinical, demographic, and longitudinal data, was provided to the AI algorithms.

In a recent, comprehensive study, the performance of a stand-alone AI algorithm for detection of breast cancer in DM images was evaluated both on large retrospective screening datasets and against the performance of readers during an observer study [74]. In the first place, the performance of the algorithm was compared to a dataset of 25,856 DM cases from the UK screening program. This involved the comparison against the two human reads of every case, and, when performed, against the final decision involving a third reader. The AI algorithm performed better than the first reader (improvement in sensitivity of +2.70 % and in specificity of +1.18 %), and non-inferiorly when compared to the performance of the second reader and to the final decision. It should be noted that in the UK screening program, the double

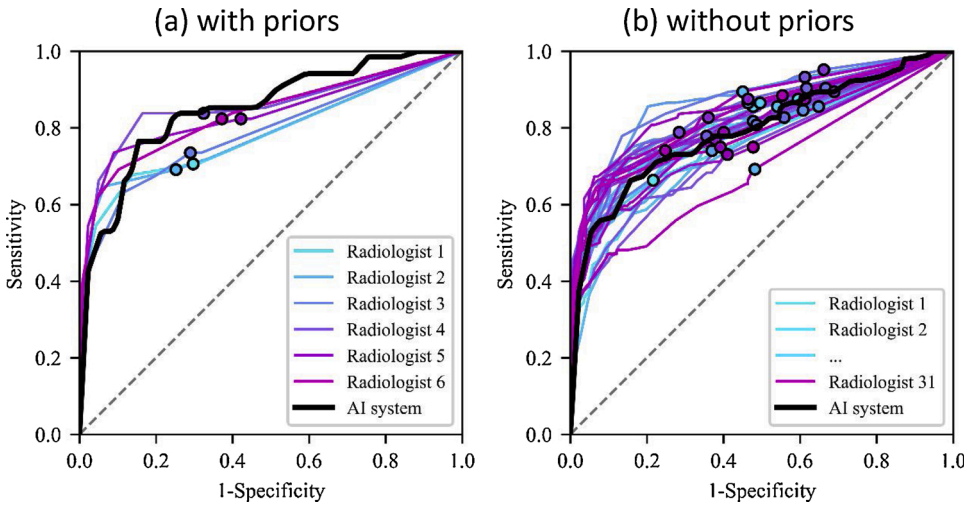

Fig. 6. Stand-alone ROC performance of a commercial AI system compared to the radiologists interpreting mammographic data sets (a) with, and (b) without the use of the prior images for comparison. Reprinted with permission from Rodriguez Ruiz et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019;111(9):916–922.

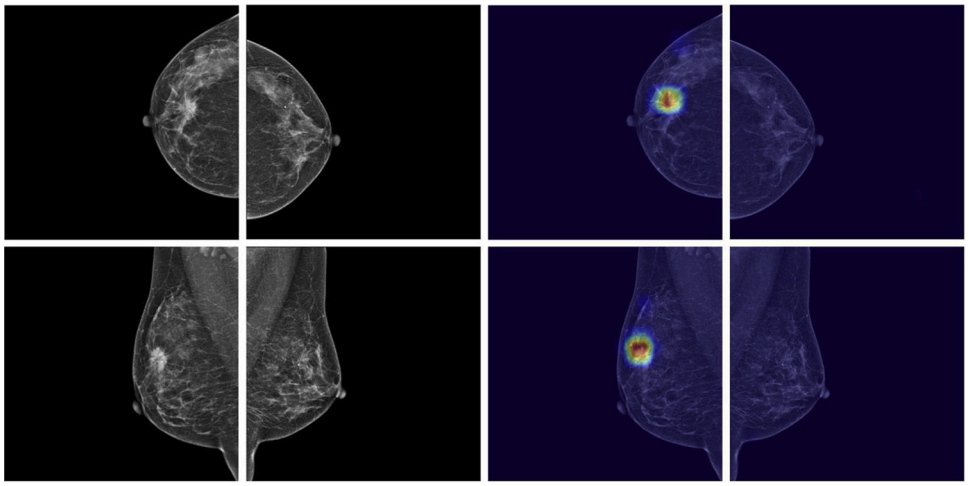

Fig. 7. (left) Digital mammography of a 44-year old woman with invasive ductal carcinoma in the right breast, with (right) an overlaid heat map highlighting the area that most strongly contributed to the final classification decision. Reprinted with permission from Kim et al. Applying Data-driven Imaging Biomarker in Mammography for Breast Cancer Screening: Preliminary Study. Sci Rep. 2018;8(1):1–8.

reading is not performed independently, and therefore the decision of the second reader probably influenced, presumably positively, by the interpretation of the first reader. The final decision is, of course, based on the opinion of a third reader with access to the first two decisions, so this performance is further enhanced. In the same study, the investigators also compared the AI performance to that of USA-based single reading of 3097 screening DM. Again, the AI outperformed the human readers in both sensitivity (+9.40 %) and specificity (+5.70 %). Interestingly, with the AI system trained on UK-only data, it still resulted in better performance on the US dataset than the US radiologists. Finally, in a comparison of the performance of six readers interpreting an enriched data set of 465 DM cases, the AI (AUC = 0.740) outperformed all six readers (average reader AUC = 0.625). Aside from the current DM images themselves, this AI algorithm is only input the age of the screened woman, also being unable to take advantage of any additional information that may result from the prior screening images.

Finally, in another study, Kim et al. evaluated an AI algorithm using three different DM datasets, from South Korea, USA, and UK. Although the performance of the AI, in terms of AUC, was high for all three datasets, the performance by radiologists on these same datasets is not reported, so it is not possible to determine the significance of these results [75]. However, the authors also performed a reader study to directly compare the performance of the AI algorithm in cancer detection in DM to that of 14 radiologists, interpreting 320 DM cases. The AI was more accurate (AUC = 0.940) than all the radiologists, and, of course, the average radiologist (AUC = 0.810).

10. AI for breast cancer detection in digital mammography and digital breast tomosynthesis: Clinical implementation options

10.1. Decision support

The commercial AI algorithm for DM described above [70], was first tested by Rodriguez Ruiz et al. for use in decision support [76]. When used in this way, the AI algorithm is not used either as a first or second reader, but it is used concurrently by the radiologist during his/her interpretation of the mammographic exam. Specifically, once the radiologist identifies a finding of concern and is deciding if it is suspicious enough to be recalled or not, the opinion of the AI may be requested when he/she cannot arrive at a decision. This is usually done by the

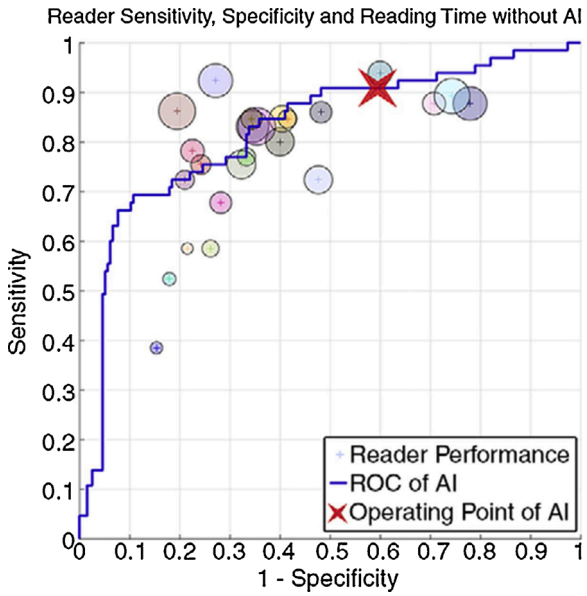

Fig. 8. Stand-alone ROC performance of an AI system for digital breast tomosynthesis compared to that of 24 radiologists, interpreting 260 cases. The size of the reader performance points represents the reading time. Reprinted with permission from Conant et al. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. Radiology: Artificial Intelligence. 2019;1(4):e180096.

reader by clicking in the area of the lesion. The AI then acts as a CADx algorithm, providing a PoM or level-of-suspicion (LoS) score. In other words, instead of the radiologist asking his/her colleague at the next review station "come over and look at this, what do you think?", he/she asks the computer for a second opinion about a specific finding. In addition to this finding-specific score, the AI algorithm also provide a LoS score for the entire case. In this study 14 radiologists reviewed 240 DM cases with and without the use of the decision support algorithm. The AUC was found to be slightly, but statistically significant, higher with AI support than without. Both sensitivity and specificity increased with the AI support, although only the former reached a significant improvement, while the overall reading time for the entire dataset used in the study was similar. However, interestingly, if the reading time differences determined during the study for the actual negative and positive cases are taken separately and used to estimate the reading time in a case set at the expected disease prevalence level of general screening, the AI-assisted screening would be about 5% faster. It seems that, thanks to the AI algorithm, radiologists spend more time evaluating the cancer cases and less time on the normal cases. Something that, intuitively, seems like a positive development.

As part of the reader study by Kim et al. described above, the authors also evaluated how the performance of the radiologists changed if they were allowed to change their interpretation decisions after seeing the output of the AI algorithm [75]. In this AI-aided interpretation, the radiologists improved substantially (AUC = 0.881) compared to their original, un-aided reading (AUC = 0.810). However, it should be noted that this performance is still lower than that of the AI alone, by a substantial margin.

10.2. AI-assisted digital breast tomosynthesis reading

As mentioned above, one of the main limitations of DBT is its lengthening of the reading time, roughly doubling the reading time compared to DM. This is one of the major reasons why DBT has not yet been embraced in large screening programs like the national or regional screening programs that exist throughout Europe. In an effort to address this issue, commercial AI products have been developed to reduce the reading time while maintaining, or, if possible, increasing reader performance. For this, these commercial products aim to ease lesion detection throughout the DBT slice stack, expecting that the interpreting radiologist will only evaluate the lesions detected by the AI and that therefore are either included in an AI-generated synthetic 2D image or are highlighted in the DBT slice stack. To ease navigation, markers may be shown on a (synthetic) 2D image, jumping to the relevant DBT slice when clicked.

The performance of such a commercial AI product for DBT, albeit one that works only on soft tissue lesions, was investigated by Benedikt et al. [77]. With this product, the synthetic 2D image is generated with the aid of this AI algorithm, which increases the visibility of soft tissue lesions in the composite image. During interpretation, when the reader clicks on such a highlighted lesion in the synthetic image, he/she is taken to the corresponding slice in the DBT stack. By performing a study with 20 radiologists interpreting 240 DBT cases twice, once with and once without the use of this AI software, the authors determined that its use resulted in a reduction in the reading time of 29 %, while the mean AUC across readers was 0.850 with the AI and 0.841 without it.

In a similar evaluation of a follow-up product from the same company, Conant et al. found similar results [72]. Specifically, the software evaluated here, already mentioned above due to its evaluation in reading DBT images in stand-alone fashion, is now capable of detecting both soft tissue lesions and calcifications, and works in a similar fashion, aiming to reduce the reading time while improving, or at least matching, the performance without it. In this study, this time performed with 24 radiologists evaluating 260 DBT cases, the investigators found a 53 % decrease in reading time, while the mean AUC increased from 0.795 to 0.852 with the addition of AI. Furthermore, both sensitivity and specificity increased significantly with the use of the AI system.

Chae et al. tested a non-commercial algorithm for this same clinical application, also finding a decrease in reading time with a concurrent maintenance of performance in reading of DBT with the use of AI during interpretation of DBT images [78]. However, the reduction in time reported by these authors is much more modest; on the order of 14 %, being similar when the readers are novices or experienced.

It should be noted that the use of AI-driven reading of DBT cases in the way described in these studies, in which radiologists will not perform a lesion search, or at least not an exhaustive one, is a change from current practice. Under this reading strategy, the findings that the interpreting radiologist will evaluate best are the ones that the AI has found and is pointing out. Therefore, the sensitivity of screening will be mostly limited to that of the AI system. In the larger studies mentioned above, by Benedikt et al. and Conant et al. [72,77], the sensitivity for detection of breast cancer increased with the use of the AI-assisted reading compared to without, while for the smaller Chae et al. [78] study no significant difference was detected, so this approach does seem promising. However, the change in the role of the computer system, and that of the interpreting radiologist needs to be well-recognized, and its potential medico-legal implications, if any, evaluated. Since all cases are, for practical purposes, reviewed by a breast radiologist, it is not clear if responsibility or liability issues would, or need to, be re-assessed. However, the reality is that the search for lesions would be driven by the AI algorithm alone, and therefore, the interpreting radiologist will have to trust the performance of the AI algorithm, and take responsibility for it.

10.3. Pre-selection of normal cases

As mentioned previously, breast cancer prevalence in the screening population is <1% [41,42]. Therefore, the vast majority of screening exams to be reviewed are normal. An AI algorithm that can operate at a very high sensitivity, and therefore with a high negative predictive value, could be used to automatically identify normal cases that do not

need to be read by a breast radiologist at all. This pre-selection would increase the efficiency of screening, having various advantages: (i) allow screening radiologists to spend more time on the suspicious cases, (ii) ameliorate the screening radiologist shortage that some countries are or will face in the near future [79], and (iii) ease the introduction of DBT, which doubles the reading time, in national/regional screening programs.

Rodriguez Ruiz et al. tested the potential impact of pre-selection on screening performance using the same algorithm previously tested against 101 radiologists [80]. Instead of the lesion-based likelihood of malignancy scores used to test its stand-alone performance, in this study the case-level PoM provided by the AI system with scores ranging from 1 to 10 (10 representing the highest malignancy present probability) was used. The AI algorithm is normalized so that in a screening case set, the cases will be distributed evenly among all ten scores. Using the same aggregated data set (>2 600 cases, >28 000 interpretations), the authors defined the number of cases that would be pre-identified as normal by the AI as all those that received a case-based PoM by the AI system below a threshold score. They then determined what the resulting AUC, number of non-human-read cancer cases, and workload reduction would be. Except for the extreme case of setting a PoM score of 9 as the threshold at and below which all cases would be pre-identified as normal (leaving only the 10 % most suspicious cases for human review), the resulting AUC was found to not be significantly reduced due to the pre-selection strategy. In a scenario with a PoM score threshold of 5, which would result in approximately halving the number of cases to be human-read, the investigators found that only 7% of cancer cases would be marked as normal by the AI system. In a more conservative scenario in which the threshold score is set to 2, and therefore the workload reduced by approximately 20 %, only 1% of cancers would be lost to pre-selection.

Using the same AI system, Lång et al. [81] evaluated the impact of introducing this pre-selection strategy on a subset of almost 10 000 cases from the Malmö Breast Tomosynthesis Screening Trial [11,26]. Setting the score threshold of 5 for denoting cases as normal, the authors determined that a reduction of about 50 % in the workload would be achieved, resulting in a mis-labeling of 8 (11.3 %) cancer cases as normal, but also reducing the false positives by 27.8 %. Three breast radiologists assessed these 8 cancers missed by the AI system as clearly visible. A more conservative threshold of 2 would result in a reduction of about 19 % in the cases needing to be human-read, while mis-labeling only a single cancer case but reducing the false positives by 5.4 %.

The same pre-selection strategy was subsequently investigated by Yala et al., using another deep learning-based system specifically developed to classify screening DM cases as *cancer free*, and therefore to not be human-interpreted, or not [82]. After training this system, the authors investigated its impact on screening outcomes retrospectively, as done by Rodriguez Ruiz et al, in this case using the results of the original interpretation during screening for the cases marked as needing to be human-read (Rodriguez Ruiz et al. used the PoM scores from the previously-performed observer studies). The authors found that the pre-selection strategy with this AI system would result in a workload reduction of almost 20 %, while maintaining sensitivity and increasing specificity from 93.5%–94.2%, which is in line with the findings of Rodriguez-Ruiz et al. and Lång et al. using a threshold score of 2.

Of course, since these studies were performed retrospectively, they assume that the radiologists' performance would not be affected by the introduction of pre-selection. This could very well not be the case. Presumably, the radiologists would know that such a pre-selection strategy is in place in the screening program, and therefore they would know that if they are reading a case it is only because an AI program graded that case as suspicious (at least above some preset threshold). Would the performance or the operating point of the radiologists change due to that knowledge? Should the radiologists be informed what the AI score for the case they are about to interpret be? Should the location of the potential finding that caused that score be pointed out? Would that lead to satisfaction-of-search and therefore real lesions being missed [83,84]? These are all questions that need to be investigated further before such a strategy can be put in place. One major driver here, of course, will eventually also be the cost-effectiveness of human versus AI evaluation.

Meanwhile, the medico-legal and ethical aspects of having only computer systems interpret certain medical images also need to be addressed. Based upon their current performance, it is certain that setting any threshold for pre-selection will lead to missed cancers by the AI system alone. At the same time, the roughly equal performance achieved with this pre-selection strategy indicates that the total number of cancers detected will remain the same (when validated in real screening populations); rather different cancers will be missed by the computer than by the breast radiologist. Whether this is problematic from a medical point of view is mainly determined by the type of breast cancers detected by both systems, which still remains to be evaluated. From a medico-legal and ethical aspects, the question of responsibility comes into play, and likely requires the design of quality control protocols for AI algorithms, as well as regular auditing of their performance by breast imaging specialists. It should, however, be noted that in other medical specialties, such as clinical chemistry or hematology, the use of a computer as a standalone interpreter of diagnostic tests became standard practice years ago.

Finally, it should be mentioned that other applications for AI in DM and DBT screening are being investigated, and surely will have an impact on the early detection of breast cancer in the future [85]. A major field that would appear to be a very suitable application for this computer technology is evaluation of DM/DBT images for breast cancer development risk, which has shown promising results in various studies [86–88]. AI methods have also been used to determine the risk of lesions being hidden by superposition of normal tissue texture in screening DM, which would prompt the use of alternative screening methods [89].

11. Conclusions

Computer systems for the detection of breast cancer in screening images, be it digital mammography and digital breast tomosynthesis images, can be expected to gain a significant impact in the future. This thinking was, of course, also broadly expressed 15 to 20 years ago about computer-aided detection and diagnosis methods, and the reality eventually showed otherwise. However, it would seem that the performance of these new systems is significantly improved compared to that of those conventional algorithms, resulting in the needed reduction of the number of false positive marks per image. Given the limitations of the current stand-alone performance evaluation studies of these new algorithms, it is still not clear how their performance compares to that of breast screening radiologists in the real screening realm, which can only be evaluated during large-scale screening trials. Once this stand-alone performance is determined, the optimal use scenario or scenarios would then need to also be investigated prospectively, again, in the actual screening setting. Only then will the potential and real-world impact of this new generation of image interpretation methods be known. Of course, beyond the completion of the technical/clinical performance assessment, depending on the use scenario, potentially a number of medico-legal and ethical issues would need to be addressed and clarified. Once all these hurdles are surpassed, it may be expected that AI will change how screening for breast cancer is performed.

Declaration of Competing Interest

All three co-authors work at the Department of Radiology and Nuclear Medicine of Radboud University Medical Center, from which an AI technology company, ScreenPoint Medical, was spun off. The Department receives royalties from this start-up company.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.

[2] E. Paci, M. Broeders, S. Hofvind, D. Puliti, S.W. Duffy, European breast Cancer service screening outcomes: a first balance sheet of the benefits and harms, Cancer Epidemiol. Biomark. Prev. 23 (7) (2014) 1159–1163.

[3] J.H. Sumkin, B.L. Holbert, J.S. Herrmann, et al., Optimal reference mammography: a comparison of mammograms obtained 1 and 2 years before the present examination, Am. J. Roentgenol. 180 (2) (2003) 343–346.

[4] M.G. Thurfjell, B. Vitak, E. Azavedo, G. Svane, E. Thurfjell, Effect on sensitivity and specificity of mammography screening with or without comparison of old mammograms, Acta radiol. 41 (1) (2000) 52–56.

[5] A.A.J. Roelofs, N. Karssemeijer, N. Wedekind, et al., Importance of comparison of current and prior mammograms in breast Cancer screening, Radiology 242 (1) (2007) 70–77.

[6] C. Varela, N. Karssemeijer, J.H.C.L. Hendriks, R. Holland, Use of prior mammograms in the classification of benign and malignant masses, Eur. J. Radiol. 56 (2) (2005) 248–255.

[7] C.M. Hakim, V.J. Catullo, D.M. Chough, et al., Effect of the availability of prior full-field digital mammography and digital breast tomosynthesis images on the interpretation of mammograms, Radiology 276 (1) (2015) 65–72.

[8] M. Posso, T. Puig, M. Carles, M. Rué, C. Canelo-Aybar, X. Bonfill, Effectiveness and cost-effectiveness of double reading in digital mammography screening: a systematic review and meta-analysis, Eur. J. Radiol. 96 (2017) 40–49.

[9] N.A. Healy, A. O'Brien, M. Knox, et al., Consensus review of discordant imaging findings after the introduction of digital screening mammography: irish national breast Cancer Screening program experience, Radiology. (2020) 181454.

[10] National Evaluation Team for Breast Cancer Screening in the Netherlands (NETB). NETB Monitor 2014 - Nation-wide Breast Cancer Screening in the Netherlands, Results 2004 -2014, Erasmus MC and Radboudumc, Rotterdam, 2019. Jan.

[11] S. Zackrisson, K. Lång, A. Rosso, et al., One-view breast tomosynthesis versus two-view mammography in the Malmö Breast Tomosynthesis Screening Trial (MBTST): a prospective, population-based, diagnostic accuracy study, Lancet Oncol. 19 (11) (2018) 1493–1503.

[12] NCI-funded Breast Cancer Surveillance Consortium Co-operative Agreement (U01CA63740). Benchmarks for Abnormal Screening Mammography Interpretations, Based on Bcsc Data, 2007 – 2013. Benchmarks for Abnormal Screening Mammography Interpretations, Based on BCSC Data, 2007 – 2013, 2017. https://www.bcsc-research.org/statistics/screening-performance-benchmarks/abnormal-scrn-benchmarks.

[13] H. Bosmans, A. De Hauwere, K. Lemmens, et al., Technical and clinical breast cancer screening performance indicators for computed radiography versus direct digital radiography, Eur. Radiol. 23 (10) (2013) 2891–2898.

[14] B. Séradour, P. Heid, J. Estève, Comparison of direct digital mammography, computed radiography, and film-screen in the french national breast Cancer Screening program, Am. J. Roentgenol. 202 (1) (2013) 229–236.

[15] I. Thomassin-Naggara, C. Balleyguier, L. Ceugnart, et al., Artificial intelligence and breast screening: French Radiology Community position paper, Diagn. Interv. Imaging 100 (10) (2019) 553–566.

[16] E.D. Pisano, C. Gatsonis, E. Hendrick, et al., Diagnostic performance of digital versus film mammography for breast-cancer screening, N. Engl. J. Med. 353 (2005) 1–11.

[17] P. Skaane, K. Young, A. Skjennald, Population-based mammography screening: comparison of screen-film and full-field digital mammography with soft-copy reading–Oslo I study, Radiology 229 (3) (2003) 877–884.

[18] P. Skaane, A. Skjennald, Screen-film mammography versus full-field digital mammography with soft-copy reading: randomized trial in a population-based screening program—the Oslo II study, Radiology 232 (1) (2004) 197–204.

[19] P. Skaane, S. Hofvind, A. Skjennald, Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: follow-up and final results of Oslo II study, Radiology 244 (3) (2007) 708–717.

[20] P. Skaane, Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review, Acta radiol. 50 (1) (2009) 3–14.

[21] NCI-funded Breast Cancer Surveillance Consortium Co-operative Agreement. Performance Measures for 1,960,150 Screening Mammography Examinations From 2002 to 2006 by Age — Based on BCSC Data As of 2009, 2011.

[22] R.L. Birdwell, D.M. Ikeda, K.F. O'Shaughnessy, E.A. Sickles, Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection, Radiology. 219 (1) (2001) 192–202.

[23] L.T. Niklason, B.T. Christian, L.T.E. Niklason, et al., Digital tomosynthesis in breast imaging, Radiology 205 (2) (1997) 399–406.

[24] S. Ciatto, N. Houssami, D. Bernardi, et al., Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study, Lancet Oncol. 14 (7) (2013) 583–589.

[25] D. Bernardi, P. Macaskill, M. Pellegrini, et al., Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study, Lancet Oncol. 17 (8) (2016) 1105–1113.

[26] K. Lång, I. Andersson, A. Rosso, A. Tingberg, P. Timberg, S. Zackrisson, Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study, Eur. Radiol. 26 (1) (2016) 184–190.

[27] F.J. Gilbert, L. Tucker, M.G.C. Gillan, et al., Accuracy of digital breast tomosynthesis for depicting breast Cancer subgroups in a UK retrospective reading study (TOMMY trial), Radiology. 277 (3) (2015) 697–706.

[28] P. Skaane, A.I. Bandos, R. Gullien, et al., Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program, Radiology. 267 (1) (2013) 47–56.

[29] S. Hofvind, T. Hovda, A.S. Holen, et al., Digital breast tomosynthesis and synthetic 2D mammography versus digital mammography: evaluation in a population-based screening program, Radiology 287 (3) (2018) 787–794.

[30] K. Ganesan, U.R. Acharya, C.K. Chua, L.C. Min, K.T. Abraham, K.-H. Ng, Computer-aided breast Cancer detection using mammograms: a review, IEEE Rev. Biomed. Eng. 6 (2013) 77–98.

[31] S. van Engeland, N. Karssemeijer, Combining two mammographic projections in a computer aided mass detection method, Med. Phys. 34 (3) (2007) 898–905.

[32] D. Tahmoush, H. Samet, Image similarity and asymmetry to improve computer-aided detection of breast cancer, in: S.M. Astley, M. Brady, C. Rose, R. Zwiggelaar (Eds.), Proceedings of International Workshop of Digital Mammography, Springer, Berlin, Heidelberg, 2006, pp. 221–228.

[33] D. Tahmoush, H. Samet, An improved asymmetry measure to detect breast cancer. Medical imaging 2007: computer-aided diagnosis, Proceedings of SPIE 6514 (2007) 65141Q.

[34] X. Wang, L. Li, W. Xu, W. Liu, D. Lederman, B. Zheng, Improving performance of computer-aided detection of masses by incorporating bilateral mammographic density asymmetry: an assessment, Acad. Radiol. 19 (3) (2012) 303–310.

[35] R.A. Castellino, Computer aided detection (CAD): an overview, Cancer Imaging 5 (1) (2005) 17–19.

[36] L.J. Warren Burhenne, S.A. Wood, C.J. D'Orsi, et al., Potential contribution of computer-aided detection to the sensitivity of screening mammography, Radiology. 215 (2) (2000) 554–562.

[37] T.W. Freer, M.J. Ulissey, Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center, Radiology 220 (3) (2001) 781–786.

[38] S.V. Destounis, P. DiNitto, W. Logan-Young, E. Bonaccio, M.L. Zuley, K.M. Willison, Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience, Radiology. 232 (2) (2004) 578–584.

[39] J.J. Fenton, S.H. Taplin, P.A. Carney, et al., Influence of computer-aided detection on performance of screening mammography, N. Engl. J. Med. 356 (14) (2007) 1399–1409.

[40] C.D. Lehman, R.D. Wellman, D.S.M. Buist, K. Kerlikowske, A.N.A. Tosteson, D. L. Miglioretti, Diagnostic accuracy of digital screening mammography with and without computer-aided detection, JAMA Intern. Med. 175 (11) (2015) 1828–1837.

[41] NCI-funded Breast Cancer Surveillance Consortium Co-operative Agreement (U01CA63740 U. Screening Mammography Sensitivity, Specificity, & False Negative Rate, BCSC, 2019. Accessed October 27, 2019, https://www.bcsc-research.org/statistics/screening-performance-benchmarks/screening-sens-spec-false-negative.

[42] National Evaluation Team for Breast Cancer Screening in the Netherlands (NETB). NETB Monitor 2013 - Nation-wide Breast Cancer Screening in the Netherlands, Results 1990-2013, Erasmus MC and Radboudumc, Rotterdam, 2015. Jul.

[43] D.M. Ikeda, R.L. Birdwell, K.F. O'Shaughnessy, E.A. Sickles, R.J. Brenner, Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography, Radiology. 230 (3) (2004) 811–819.

[44] R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, Analysis of computer-aided detection techniques and signal characteristics for clustered microcalcifications on digital mammography and digital breast tomosynthesis, Phys. Med. Biol. 61 (19) (2016) 7092–7112.

[45] J. Katzen, K. Dodelzon, A review of computer aided detection in mammography, Clin. Imaging 52 (2018) 305–309.

[46] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[47] ImageNet Large Scale Visual Recognition Competition, ILSVRC2012, 2012. Accessed October 27, 2019, http://www.image-net.org/challenges/LSVRC/2012/results.html.

[48] G. Litjens, T. Kooi, B.E. Bejnordi, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[49] W. Lotter, G. Sorensen, D. Cox, et al., A multi-scale CNN and curriculum learning strategy for mammogram classification, in: M.J. Cardoso, T. Arbel, G. Carneiro (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Cham: Springer International Publishing, 2017, pp. 169–177.

[50] E.-K. Kim, E.-K. Kim, K. Han, et al., Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study, Sci. Rep. 8 (1) (2018) 1–8.

[51] T. Kooi, G. Litjens, B. van Ginneken, et al., Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.

[52] R.K. Samala, H.-P. Chan, L.M. Hadjiiski, K. Cha, M.A. Helvie, Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis, Proceedings of SPIE 9785 (2016) 97850.

[53] M.A. Al-masni, M.A. Al-antari, J.-M. Park, et al., Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system, Comput. Methods Programs Biomed. 157 (2018) 85–94.

[54] T. Buelow, H.S. Heese, R. Grewer, D. Kutra, R. Wiemker, Inter- and intra-observer variations in the delineation of lesions in mammograms, Proceedings of SPIE 9416 (2015) 941605.
[55] H.-C. Shin, H.R. Roth, M. Gao, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298.
[56] R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, K.H. Cha, C.D. Richter, Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms, Phys. Med. Biol. 62 (23) (2017) 8894–8908.
[57] B.Q. Huynh, H. Li, M.L. Giger, Digital mammographic tumor classification using transfer learning from deep convolutional neural networks, JMI. 3 (3) (2016) 034501.
[58] S.S. Aboutalib, A.A. Mohamed, W.A. Berg, M.L. Zuley, J.H. Sumkin, S. Wu, Deep learning to distinguish recalled but benign mammography images in breast Cancer screening, Clin. Cancer Res. 24 (23) (2018) 5902–5909.
[59] R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, C. Richter, K. Cha, Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis, Phys. Med. Biol. 63 (9) (2018) 095005.
[60] D.H. Kim, S.T. Kim, Y.M. Ro, Latent feature representation with 3-D multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2016) 927–931.
[61] T. Kooi, N. Karssemeijer, Deep learning of symmetrical discrepancies for computer-aided detection of mammographic masses, Proceedings of SPIE 10134 (2017) 101341J.
[62] T. Kooi, N. Karssemeijer, Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks, JMI. 4 (4) (2017) 044501.
[63] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, R. Zwiggelaar, Deep learning in mammography and breast histology, an overview and future trends, Med. Image Anal. 47 (2018) 45–67.
[64] X. Zhang, Y. Zhang, E.Y. Han, et al., Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks, IEEE Trans. Nanobioscience 17 (3) (2018) 237–242.
[65] 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.
[66] G. van Schie, M.G. Wallis, K. Leifland, M. Danielsson, N. Karssemeijer, Mass detection in reconstructed digital breast tomosynthesis volumes with a computer-aided detection system trained on 2D mammograms, Med. Phys. 40 (4) (2013) 041902.
[67] K. Mendel, H. Li, D. Sheth, M. Giger, Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography, Acad. Radiol. 26 (6) (2019) 735–743.
[68] S.V. Fotin, Y. Yin, H. Haldankar, J.W. Hoffmeister, S. Periaswamy, Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches, Proceedings of SPIE 9785 (2016) 97850X.
[69] A.S. Becker, M. Marcon, S. Ghafoor, M.C. Wurnig, T. Frauenfelder, A. Boss, Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast Cancer, Invest. Radiol. 52 (7) (2017) 434–440.
[70] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, et al., Stand-alone artificial intelligence for breast Cancer detection in mammography: comparison with 101 radiologists, J. Natl. Cancer Inst. 111 (9) (2019) 916–922.
[71] D. Gur, A.I. Bandos, C.S. Cohen, et al., The "Laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations, Radiology 249 (1) (2008) 47–53.
[72] E.F. Conant, A.Y. Toledano, S. Periaswamy, et al., Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis, Radiology: Artificial Intelligence. 1 (4) (2019) e180096.
[73] T. Schaffter, D.S.M. Buist, C.I. Lee, et al., Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms, JAMA Netw Open. American Medical Association 3 (3) (2020) e200265–e200265.
[74] S.M. McKinney, M. Sieniek, V. Godbole, et al., International evaluation of an AI system for breast cancer screening, Nature. 577 (7788) (2020) 89–94.
[75] H.-E. Kim, H.-E. Kim, B.-K. Han, et al., 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.
[76] A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, et al., Detection of breast cancer with mammography: effect of an artificial intelligence support, Radiology. 290 (2) (2018) 305–314.
[77] R.A. Benedikt, J.E. Boatsman, C.A. Swann, A.D. Kirkpatrick, A.Y. Toledano, Concurrent computer-aided detection improves reading time of digital breast tomosynthesis and maintains interpretation performance in a multireader multicase study, Am. J. Roentgenol. 210 (3) (2017) 685–694.
[78] E.Y. Chae, H.H. Kim, J. Jeong, S.-H. Chae, S. Lee, Y.-W. Choi, Decrease in interpretation time for both novice and experienced readers using a concurrent computer-aided detection system for digital breast tomosynthesis, Eur. Radiol. 29 (5) (2019) 2518–2525.
[79] The Breast Imaging and Diagnostic Workforce in the United Kingdom | the Royal College of Radiologists, 2019. Accessed April 30, 2019, https://www.rcr.ac.uk/publication/breast-imaging-and-diagnostic-workforce-united-kingdom.
[80] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, et al., Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol. 29 (9) (2019) 4825–4832.
[81] K. Lång, M. Dustler, V. Dahlblom, I. Andersson, S. Zackrisson, Can Artificial Intelligence Identify Normal Mammograms in Screening? European Congress of Radiology, Vienna, Austria, 2019.
[82] A. Yala, T. Schuster, R. Miles, R. Barzilay, C. Lehman, A deep learning model to triage screening mammograms: a simulation study, Radiology. 293 (1) (2019) 38–46.
[83] K.S. Berbaum, E.A.J. Franken, D.D. Dorfman, et al., Satisfaction of search in diagnostic radiology, Invest. Radiol. 25 (2) (1990) 133.
[84] K.S. Berbaum, G.Y. El-Khoury, E.A. Franken, et al., Missed fractures resulting from satisfaction of search effect, Emerg. Radiol. 1 (5) (1994) 242–249.
[85] A. Arieno, A. Chan, S.V. Destounis, A review of the role of augmented intelligence in breast imaging: from automated breast density assessment to risk stratification, Am. J. Roentgenol. 212 (2) (2018) 259–270.
[86] A. Yala, C. Lehman, T. Schuster, T. Portnoi, R. Barzilay, A deep learning mammography-based model for improved breast Cancer risk prediction, Radiology. 292 (1) (2019) 60–66.
[87] K. Dembrower, Y. Liu, H. Azizpour, et al., Comparison of a deep learning risk score and standard mammographic density score for breast Cancer risk prediction, Radiology. 294 (2) (2019) 265–272.
[88] Y. Qu, G. Yue, C. Shang, L. Yang, R. Zwiggelaar, Q. Shen, Multi-criterion mammographic risk analysis supported with multi-label fuzzy-rough feature selection, Artif. Intell. Med. 100 (2019) 101722.
[89] T. Cleland, J.G. Mainprize, O. Alonzo-Proulx, et al., Use of convolutional neural networks to predict risk of masking by mammographic density, Proceedings of SPIE 10950 (2019) 109501X.