Article published online: 2023-10-10

Reproducibility and Explainability of Deep Learning in Mammography: A Systematic Review of Literature

Deeksha Bhalla[1] Krithika Rangarajan[1] Tany Chandra[1] Subhashis Banerjee[2] Chetan Arora[2]

[1] Department of Radiodiagnosis, All India Institute of Medical Sciences, New Delhi, India
[2] Department of Computer Science and Engineering, Indian Institute of Technology, New Delhi, India

Address for correspondence Krithika Rangarajan, MD, Room 47A IRCH, All India Institute of Medical Sciences, New Delhi 10029, India (e-mail: krithikarangarajan86@gmail.com).

Indian J Radiol Imaging 2024;34:469–487.

Abstract

Background Although abundant literature is currently available on the use of deep learning for breast cancer detection in mammography, the quality of such literature is widely variable.

Purpose To evaluate published literature on breast cancer detection in mammography for reproducibility and to ascertain best practices for model design.

Methods The PubMed and Scopus databases were searched to identify records that described the use of deep learning to detect lesions or classify images into cancer or noncancer. A modification of Quality Assessment of Diagnostic Accuracy Studies (mQUADAS-2) tool was developed for this review and was applied to the included studies. Results of reported studies (area under curve [AUC] of receiver operator curve [ROC] curve, sensitivity, specificity) were recorded.

Results A total of 12,123 records were screened, of which 107 fit the inclusion criteria. Training and test datasets, key idea behind model architecture, and results were recorded for these studies. Based on mQUADAS-2 assessment, 103 studies had high risk of bias due to nonrepresentative patient selection. Four studies were of adequate quality, of which three trained their own model, and one used a commercial network. Ensemble models were used in two of these. Common strategies used for model training included patch classifiers, image classification networks (ResNet in 67%), and object detection networks (RetinaNet in 67%). The highest reported AUC was 0.927 ± 0.008 on a screening dataset, while it reached 0.945 (0.919–0.968) on an enriched subset. Higher values of AUC (0.955) and specificity (98.5%) were reached when combined radiologist and Artificial Intelligence readings were used than either of them alone. None of the studies provided explainability beyond localization accuracy. None of the studies have studied interaction between AI and radiologist in a real world setting.

Conclusion While deep learning holds much promise in mammography interpretation, evaluation in a reproducible clinical setting and explainable networks are the need of the hour.

Keywords
- artificial intelligence
- breast cancer
- deep learning
- mammography
- neural networks
- systematic review

article published online
October 10, 2023

DOI https://doi.org/10.1055/s-0043-1775737.
ISSN 0971-3026.

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

Introduction

Computer-aided detection (CAD) techniques in mammography have a controversial history.

Traditional CAD used hand-crafted features to detect cancers on mammograms, and received Food and Drug Administration clearance way back in 1998. There was a lot of initial enthusiasm about the use of CAD in mammography, with many large studies suggesting it can improve detection of cancers.[1–4] These were however retrospective studies, performed in simulated environments. When deployed for clinical use, it was found that CAD actually reduces the accuracy for cancer detection, and increases biopsy rates.

Deep learning (DL) has made much headway in medical imaging. This is particularly true of breast imaging, where various studies have reported accuracies comparable with radiologists. Many studies have even suggested that DL may be used, not just as a second user, but also to triage mammograms without user intervention, thereby reducing the workload on the radiologist. This may be particularly valuable, given the increasing work-load and may even make way for breast screening in developing countries. However, even today, most studies are in retrospective simulated environments. A systematic review by Freeman et al[5] indicated that the clinical design of most studies is poor, and the level of evidence for conclusions drawn is low.

DL models essentially learn from the data they have trained on, and would carry forward biases in these data in an invisible, difficult-to-detect manner. It has been seen that results reported in the literature are often not reproducible in clinical settings. Wang et al[6] in their study demonstrated how performance varied widely when six different algorithms were tested on four mammography datasets, with a significant fall in accuracy on external validation. Thus, reproducibility is an essential metric when assessing for possible clinical deployment of any algorithm. In recent times, detailed check-lists such as the Medical Image Computing and Computer Assisted Interventions (MICCAI) reproducibility check-list[7] and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)[8] have been made available as a guide to authors planning and reporting such studies, to protect from lack of reproducibility. Thus, to assess for potential reproducibility of studies included in this systematic review, we checked for their adherence to such check-lists. The importance of explainability can be further understood by understanding the inverse relationship between simplicity of an algorithm and the performance. Unlike simpler algorithms which are inherently easier to understand, on the other hand, more advanced algorithms, especially multilayered DL-based algorithms, are known as a "black box," since little is known about what made the algorithm come to a particular conclusion. In health care, this explanation is essential for patient-centered counselling and ethical as well legal concerns. For models to be considered credible in the clinical setting, it is essential that it be known whether the predictions made by these models are clinically justifiable. The reader is referred to a review by Li et al for an understanding of various technical methods used for building trust-worthy, interpretable Artificial Intelligence (AI) models[9]

In this review, we attempt to assess currently available literature for reproducibility and explainability of these models, to take a measured view of the position of DL in mammography today. In addition, for studies which do have a robust clinical validation, we describe the best practices in model development and clinical design in detail, as adopted by these investigators. Some technical terms used in this review have been explained in online ►**Supplementary Table S1** for ease of the reader.

Materials and Methods

This systematic review was conducted as per the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guidelines.[10] The protocol was registered with the international prospective registry of systematic reviews (CRD42020222668).

Information Sources

A search of the Pubmed and Scopus databases was made in August 2021 by two independent reviewers. The keywords used were "deep learning" OR "artificial intelligence" AND "mammography" OR "breast" OR "breast cancer." The titles as well as abstracts of the studies were examined by reviewers for inclusion in the study. The identified articles were retrieved and manual search of bibliography was done to identify other potentially relevant studies.

Eligibility Criteria

Studies required to fulfill the following criteria to be considered for inclusion. (1) Studies reporting the development of a new DL model or validation of an existing commercially available model. (2) Application of model to the domain of either lesion detection or classification. (3) Information on training and performance of algorithm available in study; or if version and model of commercially available software have been mentioned. (4) Full text of article available in English language. The exclusion criteria were: (1) studies reporting only breast density assessment by models. (2) Studies reporting only accuracy of segmentation of regions of interest extracted by user on mammograms.

We also excluded review articles, opinions, letters to editors, and conference abstracts. Both reviewers examined the full texts of eligible articles to determine inclusion in the final analysis.

Data Extraction and Quality Assessment

A data extraction form was used to obtain relevant data from included studies. The dataset used for training and testing was recorded, along with the number of images in each subset. The task performed by the model along with its features, including the key-idea behind model training, and reported results were also recorded. Since the validation methodology and study design were different in each study,

we first performed a quality assessment by assessing the risk of bias and addressing applicability concerns. We devised a modification of the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool,[11] which was applicable to our research question, the "modified QUADAS-2" (mQUADAS-2). This was adapted from the CLAIM[8] as well as the MICCAI reproducibility checklist.[7] The modifications made to particularly fit AI assessment have been highlighted.

To ensure that studies we chose for detailed description had a robust study methodology, making their reported results reproducible and verifiable, we identified studies which reported their results on enriched datasets which were not representative of the distribution of breast cancer in the population, studies which did not have consecutive or random sampling, studies which did not perform external validation, studies where reference standard is not based on histopathology, and studies which had inappropriate exclusions (such as testing on only cancer images, and excluding normal images).

The entire mQUADAS-2 assessment tool is available in online as ►**Supplementary Table S2**. The quality assessment was performed by two independent reviewers (D.B. and T.C.). Differences in opinion were settled by a third reviewer (K.R.).

Data Analysis and Summary Measures

Among the studies that fulfilled the inclusion criteria and were deemed to be of acceptable diagnostic quality based on mQUADAS-2, we performed a detailed analysis of model training strategies and reported results of model training strategies. The analysis was focused on determining (1) whether the clinical study design allows for generalizability of results and (2) whether any attempt at explainability of results has been attempted in the model building and model analysis process. Based on these, the following analysis was performed.

- Clinical study design: we studied the suggested use of AI; whether AI was used as a standalone for triage of screening studies, as an aid to reporting radiologists, or direct comparison was made between AI systems and radiology readers of varying experience. We described the data collection process for training and validation of a model.
- Common practices in model design: the details of the model, including key concepts, model architecture, and hyperparameters, wherever mentioned were described
- Performance of AI: the metrics of reporting data, including accuracy, sensitivity, specificity, or area under the curve for a receiver operating characteristic (ROC) curve with confidence intervals were compared. Common metrics used included sensitivity in relation to number of false positives per image as per the free ROC curve (FROC) for detection tasks while area under the ROC curve for classification tasks. Studies were also analyzed to see if any explanation for results of AI is provided, such as lesion localization, explanation of missed cancers/false positives, or attempt at feature visualizations (or any other form of explainability).

Results

Literature Search and Study Selection

Using the search criteria, initially 12,123 articles were identified. After removal of duplicates and screening of abstracts, the full text of 179 articles was retrieved. Of these, 72 articles were excluded after screening the full texts. One hundred and seven articles were included in the final analysis.[12–118] The study selection process is summarized in ►**Fig. 1**.

Data Extraction (Initial Assessment)

Forty-seven studies tested their results on private datasets, either in isolation or in combination with public datasets. Sixty studies exclusively used publicly available datasets to report their results. Common public datasets used for testing the model included Breast Cancer Digital Repository (BCDR) (5 studies), INbreast (15 studies), Mammographic Image Analysis Society (MIAS) (17 studies), Digital Database for Screening Mammography (DDSM) (34 studies), and OPTIMAM (3 studies). Most private datasets used had only image-level labels; many authors used lesion-level labels provided in public datasets in addition to their private datasets. Initial approaches for network training included a combination of hand-engineered features and DL; several studies also used machine learning approaches such as support vector machine and random forest classifiers at some stage in their pipeline. More recent approaches use DL end-to-end. Common approaches include a standard classification network such as AlexNet, Residual Neural network (ResNet), and Visual Geometry group (VGG) trained on medical images. Many of the studies which reported detection accuracy used a standard object detection network, used for natural images. The most common networks used include Faster Regions with convolutional neural networks (RCNN),[13,18] You only look once (YOLO),[15,16,18,34] and RetinaNet.[21] To deal with availability of only small datasets with strong (lesion level) labels, authors have attempted patch learning[20,22,24,51] (classifiers trained on patches are used to initialize full image classifiers), and multi-instance learning.[22,24,69] To overcome shortage of data, several authors have mentioned performing data augmentation by flipping, rotation, and geometric transformation, while few authors have attempted generative adversarial network-based synthetic image generation for data augmentation. Most authors mention the use of transfer learning from natural images. Networks are commonly initialized with weights from ImageNet training. Common strategies used in improving accuracy included use of opposite view, opposite breast, use of full resolution images for training, multi-scale training, and use of patient metadata. Detection accuracy is commonly presented as an FROC curve which plots true-positive detections against false-positive marks per image.[119] Most common metric of classification accuracy was the area under the ROC curve.

Quality Assessment

Studies were assessed for risk of bias and applicability as per our mQUADAS-2 tool. The details of assessment are provided in ►**Table 1**. Overall, four[62,76,113,117] studies qualified

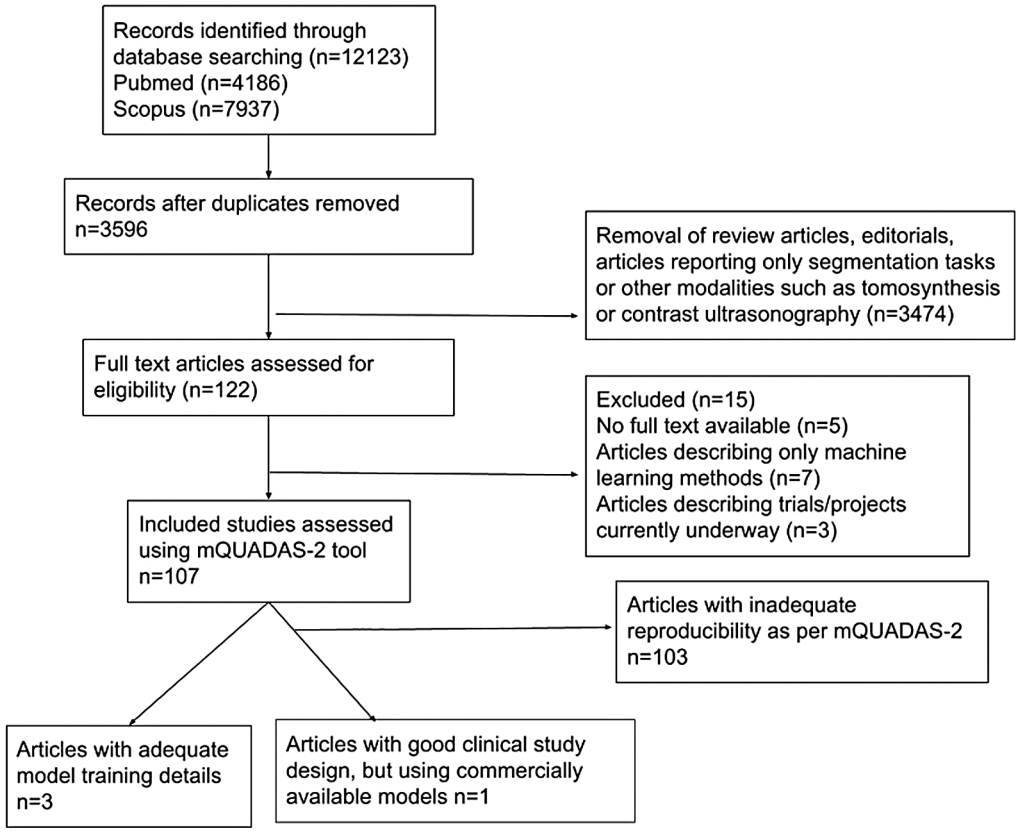

Fig. 1 Summary of study inclusion process for our review.

mQUADAS-2, of which one study described the test of a commercially available model and three described details of model training, as well as tested the model with a robust clinical study. Each of these four studies are summarized in the online ►**Supplementary Table S3**. Below, we have analyzed these studies for their clinical study design, model design and training, and their reported results.

Clinical Study Design

Studies with a robust clinical study design as ascertained by mQUADAS-2 are enlisted in ►**Table 2**. All of the studies involved retrospective patient recruitment. The study by Lång and colleagues[62] studied the utility of AI in triage of normal mammograms, while the study by Schaffter and colleagues[76] studied the performance of AI as a second reader to a radiologist. Two studies, those by Lotter and colleagues[117] and McKinney and colleagues,[113] compared the performance of AI and radiologists on similar enriched datasets. In three studies, the evaluation of AI as a standalone reader was also reported.

Training data used included the OPTIMAM dataset from the United Kingdom along with private datasets from U.S. hospitals ranging in size from 12,223 exams to 48,714 exams. The study by Schaffter et al[76] trained their model on a large dataset from the Kaiser Permanente Washington, comprising 85,580 exams which was part of the DREAM mammography challenge. Testing data size ranged from 68,026 exams from the Karolinska Institute, Sweden, which was part of the DREAM challenge, to 3,097 exams from a single institute in the United States used by McKinney et al.[113] The smallest subset used for testing was 1,533 diagnostic exams from a single institute in China used by Lotter et al.[117] In three studies, test data came from a different continent in comparison to training data. Histopathology was used for cancer proof in all studies. Length of follow-up for labeling an exam as normal or benign, ranged from 12 to 27 months. All the networks made comprehensive predictions for the entire examination, including both cranio-caudal and medio-lateral-oblique views of a patient. Image-level predictions were not made by any networks. Localization of cancer for accuracy prediction was described in two studies[113,117] while the rest of the studies did not provide any location information. Nearly all of the studies were performed in a screening setting, only the study by Lotter et al[117] tested their network on an enriched diagnostic dataset from an institute in China.

Common Practices Used for AI Model Design and Training

There were three studies which qualified mQUADAS-2 and described their models[76,113,117] (instead of using a commercial software). The three studies described eight models, which are described below.

All three studies described multi-stage pipelines,[113,117] and two of the three studies used ensembles.[76,113] All studies used lesion-level labels at some stage in their pipeline. All studies also attempted to use high resolution of images at some stage in their pipeline. The input resolution ranged from $1,100 \times 600$ to full resolution of $3,328 \times 4,096$. These

Table 1 Details of mQUADAS-2 assessment of included studies

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
1.	Deep Learning to improve breast cancer detection on screening mammography	Shen et al[38]	Aug 2019	Exclude	High	Patient selection	Only enriched datasets used
2.	Deep Learning to distinguish recalled but benign mammography images in breast cancer screening	Aboutalib et al[39]	Dec 2018	Exclude	Unclear	Patient selection	Unclear consecutive sample used or not
3.	Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer	Becker et al[40]	Jul 2017	Exclude	High	Patient selection	Split dataset
4.	Large scale deep learning for computer aided detection of mammographic lesions	Kooi et al[12]	Jan 2017	Exclude	High	Patient selection	Split dataset
5.	Discrimination of breast cancer with microcalcifications on mammography by deep learning	Wang et al[41]	Jun 2016	Exclude	High	Patient selection	Split dataset
6.	Representation learning for mammography mass lesion classification with convolutional neural networks	Arevalo et al[42]	Apr 2016	Exclude	High	Patient selection	Exclusion of normal breasts
7.	Detecting and classifying lesions in mammograms with Deep Learning	Ribli et al[13]	Mar 2018	Exclude	High	Patient selection	Only enriched datasets used
8.	Predicting breast cancer by applying deep learning to linked health records and mammograms	Akselrod-Ballin et al[114]	Aug 2019	Exclude	High	Patient selection	Split dataset
9.	A deep learning model to triage screening mammograms: a simulation study	Yala et al[43]	Oct 2019	Exclude	High	Patient selection	Split dataset
10.	A deep learning approach for the analysis of masses in mammograms with minimal user intervention	Dhungel et al[14]	Apr 2017	Exclude	High	Patient selection	Only enriched datasets used
11.	Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms	Samala et al[44]	Nov 2017	Exclude	High	Patient selection	Split dataset
12.	A deep learning-based decision support tool for precision risk assessment of breast cancer	He et al[45]	May 2019	Exclude	High	Patient selection	Test only on BIRADS 4 images
13.	Visually interpretable deep network for diagnosis of breast masses on mammograms	Kim et al[46]	Dec 2018	Exclude	High	Patient selection	Only enriched datasets used
14.	A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification	Al-Antari et al[15]	Sep 2018	Exclude	High	Patient selection	Split dataset
15.	Automated analysis of unregistered multi-view mammograms with deep learning	Carneiro et al[47]	Nov 2017	Exclude	High	Patient selection	Only enriched datasets used
16.	Deep Convolutional Neural Networks for breast cancer screening	Chougrad et al[48]	Apr 2018	Exclude	High	Patient selection	Only enriched datasets used

(*Continued*)

Table 1 (*Continued*)

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
17.	Few-shot learning with deformable convolution for multiscale lesion detection in mammography	Li et al[17]	Jul 2020	Exclude	High	Patient selection	Only enriched datasets used
18.	Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms	Cai et al[49]	Mar 2019	Exclude	High	Patient selection	Split dataset
19.	Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system	Al-Masni et al[19]	Apr 2018	Exclude	High	Patient selection	Only enriched datasets used
20.	Deep learning for mass detection in Full Field Digital Mammograms	Agarwal et al[18]	Jun 2020	Exclude	High	Patient selection	Split dataset
21.	A novel solution based on scale invariant feature transform descriptors and deep learning for the detection of suspicious regions in mammogram images	Bruno et al[50]	Jul 2020	Exclude	High	Patient selection	Only enriched datasets used
22.	Deep feature-based automatic classification of mammograms	Arora et al[51]	Jun 2020	Exclude	High	Patient selection	Only enriched datasets used
23.	Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network	Al-Masni et al[16]	Jul 2017	Exclude	High	Patient selection	Only enriched datasets used
24.	Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks	Zhang et al[52]	Jul 2018	Exclude	High	Patient selection	Cross-validation
25.	Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks	Bandeira Diniz et al[20]	Mar 2018	Exclude	High	Patient selection	Only enriched datasets used
26.	Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data	Sun et al[53]	Apr 2017	Exclude	High	Patient selection	Split dataset
27.	Improving breast mass classification by shared data with domain transformation using a generative adversarial network	Muramatsu et al[54]	Apr 2020	Exclude	High	Patient selection	Cross-validation
28.	An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization	Shen et al[55]	Dec 2020	Exclude	High	Patient selection	Split dataset
29.	Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks	Guan and Loew[56]	Jul 2019	Exclude	High	Patient selection	Split dataset of DDSM and GAN images generated from DDSM
30.	Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network	Jung et al[21]	Sep 2018	Exclude	High	Patient selection	Testing only on enriched dataset (INbreast)

Table 1 (*Continued*)

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
31.	A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets	Antropova et al[57]	Oct 2017	Exclude	High	Patient selection	Cross-validation
32.	Classification of mammogram images using multiscale all convolutional neural network (MA-CNN)	Agnes et al[58]	Dec 2019	Exclude	High	Patient selection	Testing only on enriched dataset (mini MIAS)
33.	Three-Class mammogram classification based on descriptive CNN features	Jadoon et al[59]	Jan 2017	Exclude	High	Patient selection	Cross-validation
34.	DeepCAT: deep computer-aided triage of screening mammography	Yi et al[32]	Jan 2021	Exclude	High	Patient selection	Only enriched datasets used, exclusion of microcalcification
35.	New convolutional neural network model for screening and diagnosis of mammograms	Zhang et al[60]	Aug 2020	Exclude	High	Patient selection	Testing only on enriched dataset (DDSM)
36.	Deep neural networks with region-based pooling structures for mammographic image classification	Shu et al[61]	Jun 2020	Exclude	High	Patient selection	Testing only on enriched datasets (INbreast, CBIS, DDSM)
37.	Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks	Kooi et al[63]	Oct 2017	Exclude	High	Patient selection	Split dataset
38.	Risks of feature leakage and sample size dependencies in deep feature extraction for breast mass classification	Samala et al[64]	Dec 2020	Exclude	High	Patient selection	Split dataset
39.	An ad hoc random initialization deep neural network architecture for discriminating malignant breast cancer lesions in mammographic images	Duggento et al[65]	May 2019	Exclude	High	Patient selection	Testing only on public dataset
40.	Comparison of segmentation-free and segmentation-dependent computer-aided diagnosis of breast masses on a public mammography dataset	Sawyer Lee et al[66]	Dec 2020	Exclude	High	Patient selection	Testing only on public dataset
41.	RAMS: Remote and automatic mammogram screening	Cogan et al[67]	Apr 2019	Exclude	High	Patient selection	Testing only on public dataset (INbreast)
42.	A multi-context CNN ensemble for small lesion detection	Savelli et al[22]	Mar 2020	Exclude	High	Patient selection	Only enriched dataset (INbreast), cross-validation
43.	Convolutional neural networks for the segmentation of microcalcification in mammography imaging	Valvano et al[23]	Apr 2019	Exclude	High	Patient selection	Split dataset
44.	Breast cancer detection using deep convolutional neural networks and support vector machines	Ragab et al[68]	Jan 2019	Exclude	High	Patient selection	Only enriched dataset (CBIS, DDSM)
45.	Globally-aware multiple instance classifier for breast cancer screening	Shen et al[69]	Oct 2019	Exclude	High	Patient selection	Split dataset

(*Continued*)

Table 1 (*Continued*)

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
46.	A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology	Qiu et al[70]	2017	Exclude	High	Patient selection	Cross-validation
47.	Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement	Teare et al[71]	Aug 2017	Exclude	High	Patient selection	Only enriched datasets (DDSM, ZMDS) used
48.	Detecting asymmetric patterns and localizing cancers on mammograms	Guan et al[72]	Oct 2020	Exclude	High	Patient selection	Split dataset (DREAM)
49.	Digital mammographic tumor classification using transfer learning from deep convolutional neural networks	Huynh et al[74]	Jul 2016	Exclude	High	Patient selection	Cross-validation
50.	Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning	Cha et al[26]	Jan 2020	Exclude	High	Patient selection	Only enriched datasets used (CBIS-DDSM)
51.	Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists	Rodriguez-Ruiz et al[73]	Sep 2019	Exclude	High	Patient selection	Enriched private datasets used
52.	Detection of breast cancer with mammography: effect of an artificial intelligence support system	Rodriguez-Ruiz et al[75]	Feb 2019	Exclude	High	Patient selection	Nonconsecutive sample
53.	Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms	Schaffter et al[76]	Mar 2020	Include	Low		
54.	Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women	Sasaki et al[77]	Jul 2020	Exclude	High	Patient selection	Nonconsecutive dataset
55.	Aiding the digital mammogram for detecting the breast cancer using Shearlet transform and neural network	Shenbagavalli and Thangarajan[78]	Sep 2018	Exclude	High	Patient Selection	Only enriched datasets used (DDSM)
56.	Assessing breast cancer risk with an artificial neural network	Sepandi et al[79]	Apr 2018	Exclude	High	Patient Selection	Cross-validation
57.	Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study	Rodriguez-Ruiz et al[80]	Sep 2019	Exclude	High	Patient Selection	Only enriched datasets used
58.	Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study	Kim et al[81]	Mar 2020	Exclude	High	Patient selection	Nonconsecutive sample
59.	Transfer representation learning using inception-v3 for the detection of masses in mammography	Mednikov et al[82]	Jul 2018	Exclude	High	Patient selection	Only enriched datasets (INbreast) used

Table 1 (*Continued*)

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
60.	A two-stage multiple instance learning framework for the detection of breast cancer in mammograms	Sarath et al[24]	Jul 2020	Exclude	High	Patient selection	Only enriched datasets used (INbreast)
61.	A hybridized ELM for automatic micro calcification detection in mammogram images based on multi-scale features	Melekoodappattu and Subbian[83]	May 2019	Exclude	High	Patient selection	Cross-validation, testing only on public dataset (MIAS)
62.	Applying a new quantitative image analysis scheme based on global mammographic features to assist diagnosis of breast cancer	Chen et al[84]	Oct 2019	Exclude	High	Patient selection	Cross-validation
63.	Convolutional neural networks for mammography mass lesion classification	Arevalo et al[85]	Aug 2015	Exclude	High	Patient selection	Only enriched datasets used (BCDR)
64.	Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification	Taghanaki et al[86]	Jul 2017	Exclude	High	Patient selection	Only enriched datasets used (IRMA, INbreast)
65.	Breast mass detection in digital mammogram based on Gestalt psychology	Wang et al[25]	May 2018	Exclude	High	Patient selection	Only enriched datasets used (DDSM, MIAS)
66.	A novel cascade classifier for automatic microcalcification detection	Shin et al[28]	Dec 2015	Exclude	High	Patient selection	Only enriched datasets used (MIAS, mini-MIAS)
67.	Ensemble of convolutional neural networks for classification of breast microcalcification from mammograms	Sert et al[87]	Jul 2017	Exclude	High	Patient selection	Only enriched datasets used (DDSM)
68.	A new approach to develop computer-aided detection schemes of digital mammograms	Tan et al[88]	Jun 2015	Exclude	High	Patient selection	Cross-validation
69.	A CAD system to analyze mammogram images using fully complex-valued relaxation neural network ensembled classifier	Saraswathi and Srinivasan[89]	Oct 2014	Exclude	High	Patient selection	Only enriched datasets used (MIAS)
70.	Automated breast cancer detection in digital mammograms of various densities via deep learning	Suh et al[90]	Nov 2020	Exclude	High	Patient selection	Split dataset
71.	A deep feature based framework for breast masses classification	Jiao et al[91]	Feb 2016	Exclude	High	Patient Selection	Only enriched datasets (ImageNet LSRVC, DDSM) used
72.	Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network	Kooi et al[92]	Mar 2017	Exclude	High	Patient selection	Cross-validation
73.	Global detection approach for clustered microcalcifications in mammograms using a deep learning network	Wang et al[27]	Apr 2017	Exclude	High	Patient selection	Split dataset

(*Continued*)

Table 1 (*Continued*)

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
74.	Computer-aided mammogram diagnosis system using deep learning convolutional fully complex-valued relaxation neural network classifier	Duraisamy and Emperumal[93]	Dec 2017	Exclude	High	Patient selection	Only enriched datasets (MIAS + BCDR) used
75.	Deep learning versus classical neural approach to mammogram recognition	Kurek et al[94]	Dec 2018	Exclude	High	Patient selection	Only enriched datasets (DDSM) used
76.	A parasitic metric learning net for breast mass classification based on mammography	Jiao et al[95]	Mar 2018	Exclude	High	Patient selection	Only enriched datasets used (DDSM)
77.	An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network	Al-antari et al[96]	Sep 2017	Exclude	High	Patient selection	Only enriched datasets (DDSM) used
78.	A context-sensitive deep learning approach for microcalcification detection in mammograms	Wang and Yangl[29]	June 2018	Exclude	High	Patient selection	Dataset collection method unlikely consecutive
79.	Multi-view feature fusion based four views model for mammogram classification using convolutional neural network	Nasir Khan et al[99]	Nov 2019	Exclude	High	Patient selection	Only enriched datasets (CBIS-DDSM, MIAS) used
80.	Detection of abnormalities in mammograms using deep features	Tavakoli et al[103]	Dec 2019	Exclude	High	Patient selection	Only enriched dataset (MIAS), split dataset
81.	A deep learning approach for breast cancer mass detection	Fathy and Ghoneim[30]	2019	Exclude	High	Patient selection	Only enriched dataset (DDSM), split dataset
82.	A new triplet convolutional neural network for classification of lesions on mammograms	Medjeded et al[100]	Oct 2019	Exclude	High	Patient selection	Only enriched datasets (DDSM and MIAS) used
83.	Multi-view convolutional neural networks for mammographic image classification	Sun et al[101]	Sep 2019	Exclude	High	Patient Selection	Only enriched datasets (MIAS, DDSM) used
84.	Transferring deep neural networks for the differentiation of mammographic breast lesions	Yu et al[102]	Dec 2018	Exclude	High	Patient selection	Only enriched datasets (BCDR) used
85.	Deep learning for breast cancer diagnosis from mammograms—a comparative study	Tsochatzidis et al[118]	Mar 2019	Exclude	High	Patient Selection	Only enriched datasets (CBIS-DDSM, DDSM) used
86.	Application of deep learning in the detection of breast lesions with four different breast densities	Li et al[31]	July 2021	Exclude	High	Patient selection	Enriched private dataset used for testing
87.	Breast mass detection in mammography based on image template matching and CNN	Sun et al[33]	Apr 2021	Exclude	High	Patient selection	Only enriched datasets (DDSM) used
88.	Impact of image compression on deep learning-based mammogram classification	Jo et al[104]	Apr 2021	Exclude	High	Patient Selection	Cross-validation
89.	Improving the prediction of benign or malignant breast masses using a combination of image biomarkers and clinical parameters	Cui et al[105]	Mar 2021	Exclude	High	Patient selection	• Split dataset • Exclusion of benign images

Table 1 (*Continued*)

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
90.	Compare and contrast: detecting mammographic soft-tissue lesions with C 2-Net	Liu et al[37]	Jul 2021	Exclude	High	Patient selection	Split dataset
91.	Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography	Chouhan et al[107]	May 2021	Exclude	High	Patient selection	Only enriched datasets (IRMA) used
92.	Microscopic tumour classification by digital mammography	Yang et al[106]	Feb 2021	Exclude	High	Patient selection	Split dataset
93.	A framework for breast cancer classification using Multi-DCNNs	Ragab et al[108]	Apr 2021	Exclude	High	Patient selection	Only enriched datasets (DDSM, MIAS) used
94.	Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses	Tsochatzidis et al[109]	Mar 2021	Exclude	High	Patient selection	Only enriched datasets (DDSM, CBIS-DDSM) used
95.	YOLO based breast masses detection and classification in full-field digital mammograms	Aly et al[34]	Mar 2021	Exclude	High	Patient selection	Only enriched datasets (INbreast) used
96.	Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network	Rehman et al[111]	Jul 2021	Exclude	High	Patient Selection	Only enriched datasets (DDSM, Pinum) used
97.	Presentation of novel hybrid algorithm for detection and classification of breast cancer using growth region method and probabilistic neural network	Isfahani et al[35]	Jun 2021	Exclude	Unclear	Patient selection	Only enriched datasets (DDSM, BIRADS) used
98.	Pattern classification for breast lesion on FFDM by integration of radiomics and deep features	Zhang et al[110]	Jun 2021	Exclude	Unclear	Patient selection	• Nonconsecutive sample • Split dataset
99.	Multi-scale attention-based convolutional neural network for classification of breast masses in mammograms	Niu et al[97]	Jul 2021	Exclude	High	Patient selection	Only enriched datasets (DDSM INbreast) used
100.	Mammogram mass segmentation and detection using Legendre neural network-based optimal threshold	Sarangi et al[36]	Apr 2021	Exclude	High	Patient selection	Only enriched datasets (MIAS) used
101.	Optimized radial basis neural network for classification of breast cancer images	Rajathi et al[98]	2021	Exclude	High	Patient selection	Only enriched datasets (MIAS) used
102.	External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms	Salim et al[112]	2020	Exclude	High	Patient selection	Case control design
103.	Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach	Lotter et al[117]	Feb 2021	Include	Low		
104.	Identifying normal mammograms in a large screening population using artificial intelligence	Lång et al[62]	2020	Include	Low		

(*Continued*)

Table 1 (*Continued*)

	Title	Author	Year	Decision toward detailed analysis	Risk of bias	Reason for exclusion: domain	Reason
105.	Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool	Pacilè et al[115]	2020	Exclude	High	Patient selection	Enriched private dataset used
106.	Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography	Watanabe et al[116]	2019	Exclude	High	Patient selection	Enriched private dataset used
107.	International evaluation of an AI system for breast cancer screening	McKinney et al[113]	Jan 2020	Include	Low		

Abbreviations: BCDR, Breast Cancer Digital Repository; MIAS, Mammographic Image Analysis Society; DDSM Digital Database for Screening Mammography.

were provided either through patches generated at full resolution[76,117] or as direct input of full-resolution images.[113] Only one of these three studies explicitly described use of medically relevant information from the opposite breast and opposite view.[113] Common data augmentation techniques included resizing, rotations, and vertical flipping. Two of the models[76,113] also used patient metadata such as age to attempt to improve their performance. A summary of description of the key idea in each model is given in ►**Table 3**. ►**Fig. 2** summarizes the workflow among these four studies that were analyzed in detail.

Performance of AI

Since all studies have reported performance on widely different datasets, they are not directly comparable. However, within the category of studies which were assessed as being high quality, similar methodology was used to curate the data. Therefore, these results are tabulated in ►**Table 4**.

Two studies[113,117] compared the performance of AI against a radiologist. Although both performed this analysis only on a small enriched subset of their dataset, both reported a slightly higher performance of AI in comparison to the radiologist. One study compared the performance of radiologists with and without AI, and showed that the performance of radiologists with AI is better than either the radiologist or AI alone.[76]

All studies provided localization-based explainability, though only one evaluated localization accuracy by means of mROC curves[113] (another study provided lesion detection accuracy; however, this was restricted to location in terms of laterality and quadrant[117]). This was also only in a small subset of the test population. No other form of interpretability or explainability has been attempted in any study.

Discussion

In this review we found that although a very large number of studies have been published in scientific literature on DL in mammography, a very miniscule number of these have actually tested their results in a robust clinical study. Importantly, no study offers any explainability beyond identification of lesions (either by bounding box prediction or saliency maps).

We identified four studies which tested their results in a reproducible manner,[62,76,113,117] out of which three described their in-house models. For these we also describe the practices they used for model building.

Common Practices for Model Design

All identified studies had some common features in model design. First, all of them attempted to use images with as high resolution as possible, at some stage in the network. This stresses on the importance of the fact that despite memory constraints, it is important to preserve the resolution of images while giving them as input to neural networks. This is consistent with medical knowledge on the need for exceptionally high spatial resolution for mammograms. Second, all authors stress on the importance of using precise

Table 2 Details of studies with adequate clinical design as per mQUADAS-2 tool

Author	Testing dataset	Training dataset	Task performed	Classification level	BB annotation	Model availability	Patient recruitment	Follow-up period for negative studies	Location of cancer; indicated	External validation set (country/ continent/race)	Limitation/ explainability
Lång et al[62]	Testing: screening exams 9,581 (Malmö Breast Tomosynthesis Screening Trial)	NA	Network provides a continuous score ranging between 1 and 10 representing the level of suspicion of cancer present	Patient level, including both CC and MLO views	NA	Transpara 1.4.0	Retrospective	Nil	No	Urban Swedish population	Failure analysis performed- both small and large cancers missed. 85.7% missed cancers in dense breasts.
McKinney et al[113]	UK test set: 25,856 U.S. test set 3,097	UK set: (OPTIMAM) 13,918 (train) 62,866 (tune) US set: 12,224 (training) 3,334 (tuning)	AI standalone Comparison of AI with radiologist	Patient level	Yes	Available (►**Table 3**)	Retrospective	≥21 mo	Yes	Trained on UK population, evaluated on U.S. population	Failure analysis performed: yes Explainability: only localization
Lotter et al[117]	Screening data: 2,743 (OMI-DB), 7,951 (private US) Diagnostic data: 1,533 (private China)	Screening data: OMI-DB (23,396), DDSM (2,282), Private US (48,714)	AI standalone AI vs. radiologist	Patient level	Yes	Available (►**Table 3**)	Retrospective	18 mo: private testing dataset	Yes	Trained on UK and U.S. datasets, tested on U.S. and Chinese dataset	Failure analysis performed: yes Explainability: only localization
Schaffter et al[76]	68,026 Sweden (screening examinations)	Screening examinations Private (59,923) Private: 25,657 (validation)	AI standalone AI and radiologist	Patient level	Yes	Available (►**Table 3**)	Retrospective	12 mo: KPW 18–24 mo: KI	No	Ext validation set based on Stockholm Sweden, KI	Failure analysis performed: yes Explainability: lower performance when compared with consensus radiologist interpretation, since trained with only single radiologist interpretation

Abbreviations: AI, artificial intelligence; BB, bounding box; KI, Karolinska Institute; KPW, Kaiser Permanente Washington; NA, not applicable.

Table 3 Analysis of models employed by studies with adequate clinical design

Author	Model description
Schaffter et al[76]	Ensemble, each model of the ensemble came from the top winners in a grand challenge.
	The first model, developed by Therapixel, was a modification of VGG Net. The network was modified to reduce the number of parameters, so that it could accept a larger input size of image. The team reduced the resolution of DM images to 1152 × 832 pixels. They also reduced the number of pooling layers to detect fine features. To deal with the problem of the image having a weak signal due to presence of very small object in comparison to size of image, they first pretrained with strongly labeled data (with image patches with position information). To deal with class imbalance, they trained this with minibatches containing equal number of negative and positive samples.
	The second model developed by Ribli et al was an object detection network, and predictions were used to generate classification scores. They trained a faster RCNN on public data and some hand-annotated component of the challenge data.
	The third model developed by Guan et al[72] trained multiple segmentation models (four different models) and combined the result of these four models. The models used a combination of high-resolution images with a sliding window approach for calcification detection and low-resolution images for mass detection. They also trained the model using public datasets which contained location information, like the other authors.
	The final model developed by DeepHealth consisted of two patch level classifiers (ResNet) at two different scales for microcalcifications and masses. They used these to initialize the whole image classifier with a scanning window approach.
McKinney et al[113]	Ensemble of three models, each working at a different level of interpretation of mammograms (lesion level, breast level, and case level), each model producing a breast cancer risk score between 0 and 1 for the entire patient.
	First stage of MODEL 1 was a RetinaNet object detector trained on full mammogram images rescaled to 2,048 × 2,048. Rectangular bounding boxes were produced along with a confidence score, and the top 10 boxes among all 4 views were chosen. These patches were rescaled to 409 × 409 and a corresponding patch from the opposite breast was chosen after rough registration of the breasts. Along with this, patient age, laterality, detection coordinates, and view were concatenated. This was passed through a Mobilenet architecture. A cancer score was obtained for each patch which was combined into a case level score. The second stage of MODEL 1 took these fixed size detections and trained them with a classification model that used case level labels. At train time, 5 such crops were used per case, and at test time, 10 such crops were used, and average predictions determined.
	MODEL 2 was a breast level model. Here each image after augmentation was run through a ResNet 50 feature extractor and the final feature vector obtained from all four breasts were concatenated. This concatenated feature vector was run though a few residual blocks, convolutional blocks, and then an average pool was performed to obtain a prediction score per breast.
	MODEL 3 was a case level model, this also involved a ResNet as a feature extractor from each of the four images. Data augmentation was used and input size of 2,048 × 2,048 was used. The four feature vectors were concatenated and a single hidden layer of size 512 was applied to the combined feature vector followed by a binary classification. This ResNet was initialized with trained weights of the backbone of the object detector used by MODEL 1.
Lotter et al[117]	3-stage model. In the first stage a ResNet classifier was trained on patches of 275 × 275 obtained from full mammogram images. In the first stage they performed a 5-class classification into mass, calcification, focal asymmetry, architectural distortion, or no lesion. The same classifier was further trained to give a 3-class classification as normal, benign, or malignant. In the next stage (stage 2), this trained ResNet weights were used to initialize the backbone for a RetinaNet object detector. The images for RetinaNet were resized to 1,750 pixels (other dimension modified to maintain aspect ratio). Stage 3 consisted of a multi-instance learning-based object detector trained with only image-level labels.

Abbreviation: DM, digital mammography.

location information on cancers. This is because the malignant lesion tends to occupy a very small portion of the image. Therefore, purely classification networks which work only on image-level labels tend to perform far inferior to studies which use location of cancer. Third, all networks use transfer learning from natural images and use some form of data augmentation. Fourth, medically relevant information such as metadata and information from opposite view and opposite breast adds greatly to network performance. All the above point toward the importance of core radiology knowledge in network design. While all networks provide lesion location as a means of explainability and to check saliency of network predictions, none of the networks have explicitly discussed any other means of studying explainability.

Common Practices for Clinical Design

All the identified studies were retrospective and performed in a screening environment. All networks used large datasets

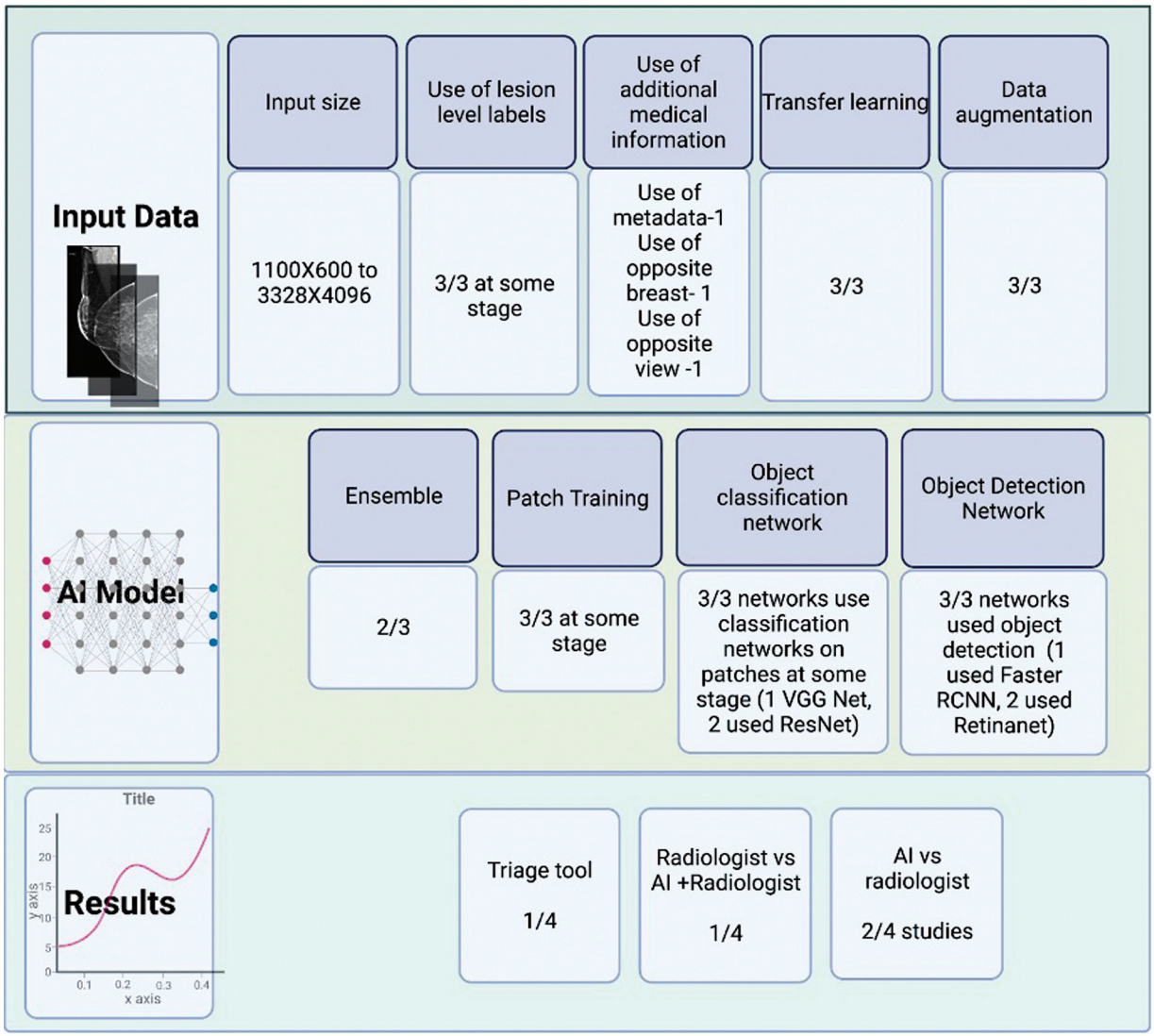

Fig. 2 Summary of detailed analysis of studies which qualified mQUADAS-2.

for training, and tested on datasets ranging from 3,000 to 68,000 mammograms[76,113]. While all the studies concluded that AI can be used for triage, or as an assistant to a radiologist as a second reader to improve accuracy, no analysis has been performed to understand the effect of false positives suggested by AI on the recall tendency of the radiologist. All studies mention the number of false-negative (missed) cancers, and some even compare the numbers with the corresponding numbers missed by radiologists in their studies. The characteristics of cancers missed by AI have also been analyzed by authors,[62,113,117] to determine patterns based on breast density, tumor size, and histological type, among others, but no consistent patterns emerged that could provide a medically sound reason for the miss. This would be of great importance in the event of potential deployment, where it would be of vital importance to explain to a patient why her cancer may have been missed by AI. In addition, among the four studies that we analyzed, only two studies mentioned confidence intervals of area under the curve for ROC curves in the results,[113,117] calling into question the possible variability in results described by the other studies. An objective measure of localization accuracy, determined by the mROC curve, was also mentioned in only a single study of these four. This is however understandable, as evaluating localization accuracy would need lesion-level labels for the entire test dataset, which would be very expensive to obtain.

Studies that report detection of interval cancers on pre-index mammograms do not mention the specificity level at which the cancer was caught on the pre-index study. Thus, how this would translate in a real-world setting remains to be seen.

Evaluation on a diagnostic mammography dataset was performed only in a single study,[117] which tested on an enriched dataset that consisted of a consecutive sample of cancers (34.8%) along with a random sample of noncancers (63.2%). Similarly, this was the only dataset from a previously unscreened population. Thus, little is known on how these networks would behave when deployed in such an environment. There were no studies that tested the AI on computed

Table 4 Reported results for various tasks performed by the AI networks

Author	AI standalone AUC	AI+ Radiologist AUC	Radiologist standalone AUC	Others
Mckinney et al[113]		NA	ROC curve encompasses average radiologist performance point 0.625 (SD 0.032)	Model sensitivity: 56.24% Model specificity: 84.29% Non-inferiority compared to radiologist
US dataset	0.757 (0.732-0.780)			
Enriched dataset (465)	0.740 (0.696-0.794)			
Lotter et al[117]		Nil	0.891 (±0.019) (best reader AUC)	Model sensitivity: 96.2% (91.7-99.2) 14.2% higher than radiologist Model specificity: 90.9% (84.9-96.1) 24% higher than radiologist
US dataset	0.927 ± 0.008			
Enriched dataset (285)	0.945 (0.919-0.968)			
Schaffter et al[76] Sweden dataset: Ensemble model	0.923	0.955 (consensus radiologist)		Specificity model: 92.5% Radiologist 96.7% (96.6-96.8) Combined model plus radiologist 98.5% (98.4-98.6)
Lang et al[62]	–	–	–	Missed cancers= 10.3% (3.1–17.5)

Abbreviations: AI, artificial intelligence; AUC, area under the curve; ROC, receiver operating curve; SD, standard deviation.

radiography systems, which are still present in many developing countries.

A recently published systematic review by Uzun Ozsahin et al[120] similarly highlights the differences and inhomogeneity in the developmental methodologies of AI algorithms but with a general sense of improvement in the quality of studies with passing time.

Overall Assessment of Position of AI in Breast Imaging

As radiology, like every other specialty in medicine and indeed every other industry, gears up for a transformation in the form of introduction of AI within the work-flow, reproducibility and explainability of neural networks form the essential building blocks of such implementation.

We found in our review that both reproducibility and explainability continue to stand in question, and would need significantly more research prior to potential clinical deployment. We thus suggest these to be important check-points for radiologists, when attempting to assess commercially available algorithms for deployment in their department. We also refer the readers to the MICCAI reproducibility checklist[7] and the CLAIM checklist[8] while designing a study to ensure their studies are of adequate quality. In our study, two algorithms performed better than radiologists at classifying mammograms; however, these had relatively small testing datasets. On the other hand, in the study with the largest testing dataset, radiologist reading showed considerably higher specificity. While it is clear that when used in the correct clinical scenario, AI holds great potential, a nuanced view should be taken to how and in what capacity it may be deployed, and where it can provide real clinical benefit.

Funding
This work was supported in part by the Department of Biotechnology, Government of India, under grant BT/PR33193/AI/133/5/2019.

Conflict of Interest
None declared.

Acknowledgment
We acknowledge the effort of our data entry operator Hema Malhothra, for her meticulous work.

References

1 Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215(02):554–562
2 Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001;219(01):192–202
3 Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. Radiology 2005;236(02):451–457
4 Brem RF, Baum J, Lechner M, et al. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. Am J Roentgenol 2003;181(03):687–693
5 Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ 2021;374:n1872
6 Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent performance of deep learning models on mammogram classification. J Am Coll Radiol 2020;17(06):796–803
7 miccai reproducibility - Google Search. Accessed May 16, 2022 at: https://www.google.com/search?q=miccai+reproducibility&oq=miccai+&aqs=chrome.1.69i57j35i39j0i512j0i20i263i512j0i512l6.4510j0j4&sourceid=chrome&ie=UTF-8

8 Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2(02):e200029
9 Li X, Xiong H, Li X, et al. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 2022;64(12):3197–3234
10 McInnes MDF, Moher D, Thombs BD, et al; and the PRISMA-DTA Group. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: the PRISMA-DTA statement. JAMA 2018;319(04):388–396
11 Whiting PF, Rutjes AWS, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(08):529–536
12 Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–312
13 Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep 2018;8(01):4165
14 Dhungel N, Carneiro G, Bradley AP. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med Image Anal 2017;37:114–128
15 Al-Antari MA, Al-Masni MA, Choi MT, Han SM, Kim TS. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. Int J Med Inform 2018;117:44–54
16 Al-Masni MA, Al-Antari MA, Park JM, et al. Detection and classification of the breast abnormalities in digital mammograms via regional Convolutional Neural Network. Annu Int Conf IEEE Eng Med Biol Soc 2017;2017:1230–1233
17 Li C, Zhang D, Tian Z, Du S, Qu Y. Few-shot learning with deformable convolution for multiscale lesion detection in mammography. Med Phys 2020;47(07):2970–2985
18 Agarwal R, Díaz O, Yap MH, Lladó X, Martí R Deep learning for mass detection in full field digital mammograms. Comput Biol Med 2020;121:103774
19 Al-Masni MA, Al-Antari MA, Park JM, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Comput Methods Programs Biomed 2018;157:85–94
20 Bandeira Diniz JO, Bandeira Diniz PH, Azevedo Valente TL, Corrêa Silva A, de Paiva AC, Gattass M. Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks. Comput Methods Programs Biomed 2018;156:191–207
21 Jung H, Kim B, Lee I, et al. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLoS One 2018;13(09):e0203355
22 Savelli B, Bria A, Molinara M, Marrocco C, Tortorella F. A multi-context CNN ensemble for small lesion detection. Artif Intell Med 2020;103:101749
23 Valvano G, Santini G, Martini N, et al. Convolutional neural networks for the segmentation of microcalcification in mammography imaging. J Healthc Eng 2019;2019:9360941
24 Sarath CK, Chakravarty A, Ghosh N, Sarkar T, Sethuraman R, Sheet D. A two-stage multiple instance learning framework for the detection of breast cancer in mammograms. Annu Int Conf IEEE Eng Med Biol Soc 2020;2020:1128–1131
25 Wang H, Feng J, Bu Q, et al. Breast mass detection in digital mammogram based on gestalt psychology. J Healthc Eng 2018; 2018:4015613
26 Cha KH, Petrick N, Pezeshk A, et al. Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning. J Med Imaging (Bellingham) 2020;7(01):012703
27 Wang J, Nishikawa RM, Yang Y. Global detection approach for clustered microcalcifications in mammograms using a deep learning network. J Med Imaging (Bellingham) 2017;4(02): 024501
28 Shin SY, Lee S, Yun ID, et al. A novel cascade classifier for automatic microcalcification detection. PLoS One 2015;10(12): e0143725
29 Wang J, Yang Y. A context-sensitive deep learning approach for microcalcification detection in mammograms. Pattern Recognit 2018;78:12–22
30 Fathy W, Ghoneim A. A deep learning approach for breast cancer mass detection. Int J Adv Comput Sci Appl 2019;10;175–182
31 Li H, Ye J, Liu H, et al. Application of deep learning in the detection of breast lesions with four different breast densities. Cancer Med 2021;10(14):4994–5000
32 Yi PH, Singh D, Harvey SC, Hager GD, Mullen LA. DeepCAT: deep computer-aided triage of screening mammography. J Digit Imaging 2021;34(01):27–35
33 Sun L, Sun H, Wang J, Wu S, Zhao Y, Xu Y. Breast mass detection in mammography based on image template matching and CNN. Sensors (Basel) 2021;21(08):2855
34 Aly GH, Marey M, El-Sayed SA, Tolba MF. YOLO based breast masses detection and classification in full-field digital mammograms. Comput Methods Programs Biomed 2021;200:105823
35 Isfahani ZN, Jannat-Dastjerdi I, Eskandari F, Ghoushchi SJ, Pourasad Y. Presentation of novel hybrid algorithm for detection and classification of breast cancer using growth region method and probabilistic neural network. Comput Intell Neurosci 2021; 2021:5863496
36 Sarangi S, Rath NP, Sahoo HK. Mammogram mass segmentation and detection using Legendre neural network-based optimal threshold. Med Biol Eng Comput 2021;59(04):947–955
37 Liu Y, Zhou C, Zhang F, et al. Compare and contrast: detecting mammographic soft-tissue lesions with C^2-Net. Med Image Anal 2021;71:101999
38 Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Sci Rep 2019;9(01):12495
39 Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin Cancer Res 2018;24 (23):5902–5909
40 Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52(07):434–440
41 Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Sci Rep 2016;6:27327
42 Arevalo J, González FA, Ramos-Pollán R, Oliveira JL, Guevara Lopez MA. Representation learning for mammography mass lesion classification with convolutional neural networks. Comput Methods Programs Biomed 2016;127:248–257
43 Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. Radiology 2019;293(01):38–46
44 Samala RK, Chan HP, Hadjiiski LM, Helvie MA, Cha KH, Richter CD. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol 2017;62(23):8894–8908
45 He T, Puppala M, Ezeana CF, et al. A deep learning-based decision support tool for precision risk assessment of breast cancer. JCO Clin Cancer Inform 2019;3:1–12
46 Kim ST, Lee JH, Lee H, Ro YM. Visually interpretable deep network for diagnosis of breast masses on mammograms. Phys Med Biol 2018;63(23):235025
47 Carneiro G, Nascimento J, Bradley AP. Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans Med Imaging 2017;36(11):2355–2365

48 Chougrad H, Zouaki H, Alheyane O. Deep convolutional neural networks for breast cancer screening. Comput Methods Programs Biomed 2018;157:19–30
49 Cai H, Huang Q, Rong W, et al. Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. Comput Math Methods Med 2019;2019:2717454
50 Bruno A, Ardizzone E, Vitabile S, Midiri M. A novel solution based on scale invariant feature transform descriptors and deep learning for the detection of suspicious regions in mammogram images. J Med Signals Sens 2020;10(03):158–173
51 Arora R, Rai PK, Raman B. Deep feature-based automatic classification of mammograms. Med Biol Eng Comput 2020;58(06): 1199–1211
52 Zhang X, Zhang Y, Han EY, et al. Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks. IEEE Trans Nanobiosci 2018;17(03):237–242
53 Sun W, Tseng TB, Zhang J, Qian W. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Comput Med Imaging Graph 2017;57:4–9
54 Muramatsu C, Nishio M, Goto T, et al. Improving breast mass classification by shared data with domain transformation using a generative adversarial network. Comput Biol Med 2020; 119:103698
55 Shen Y, Wu N, Phang J, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med Image Anal 2021;68:101908
56 Guan S, Loew M. Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. J Med Imaging (Bellingham) 2019;6(03):031411
57 Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. Med Phys 2017;44(10):5162–5171
58 Agnes SA, Anitha J, Pandian SIA, Peter JD. Classification of mammogram images using multiscale all convolutional neural network (MA-CNN). J Med Syst 2019;44(01):30
59 Jadoon MM, Zhang Q, Haq IU, Butt S, Jadoon A. Three-class mammogram classification based on descriptive CNN features. BioMed Res Int 2017;2017:3640901
60 Zhang C, Zhao J, Niu J, Li D. New convolutional neural network model for screening and diagnosis of mammograms. PLoS One 2020;15(08):e0237674
61 Shu X, Zhang L, Wang Z, Lv Q, Yi Z. Deep neural networks with region-based pooling structures for mammographic image classification. IEEE Trans Med Imaging 2020;39(06):2246–2255
62 Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol 2021;31(03): 1687–1692
63 Kooi T, Karssemeijer N. Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. J Med Imaging (Bellingham) 2017;4(04):044501
64 Samala RK, Chan HP, Hadjiiski L, Helvie MA. Risks of feature leakage and sample size dependencies in deep feature extraction for breast mass classification. Med Phys 2021;48(06):2827–2837
65 Duggento A, Aiello M, Cavaliere C, et al. An ad hoc random initialization deep neural network architecture for discriminating malignant breast cancer lesions in mammographic images. Contrast Media Mol Imaging 2019;2019:5982834
66 Sawyer Lee R, Dunnmon JA, He A, Tang S, Ré C, Rubin DL. Comparison of segmentation-free and segmentation-dependent computer-aided diagnosis of breast masses on a public mammography dataset. J Biomed Inform 2021;113:103656
67 Cogan T, Cogan M, Tamil L. RAMS: remote and automatic mammogram screening. Comput Biol Med 2019;107:18–29
68 Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ 2019;7:e6201
69 Shen Y, Wu N, Phang J, et al. Globally-aware multiple instance classifier for breast cancer screening. Mach Learn Med Imaging 2019;11861:18–26
70 Qiu Y, Yan S, Gundreddy RR, et al. A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology. J XRay Sci Technol 2017;25(05): 751–763
71 Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digit Imaging 2017;30(04):499–505
72 Guan Y, Wang X, Li H, et al. Detecting asymmetric patterns and localizing cancers on mammograms. Patterns (N Y) 2020;1(07): 100106
73 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019; 111(09):916–922
74 Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging (Bellingham) 2016;3(03):034501
75 Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290(02):305–314
76 Schaffter T, Buist DSM, Lee CI, et al; and the DM DREAM Consortium. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3(03):e200265
77 Sasaki M, Tozaki M, Rodríguez-Ruiz A, et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. Breast Cancer 2020;27(04):642–651
78 P S, R T. Aiding the digital mammogram for detecting the breast cancer using shearlet transform and neural network. Asian Pac J Cancer Prev 2018;19(09):2665–2671
79 Sepandi M, Taghdir M, Rezaianzadeh A, Rahimikazerooni S. Assessing breast cancer risk with an artificial neural network. Asian Pac J Cancer Prev 2018;19(04):1017–1019
80 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol 2019;29(09):4825–4832
81 Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020;2(03):e138–e148
82 Mednikov Y, Nehemia S, Zheng B, Benzaquen O, Lederman D. Transfer representation learning using inception-V3 for the detection of masses in mammography. Annu Int Conf IEEE Eng Med Biol Soc 2018;2018:2587–2590
83 Melekoodappattu JG, Subbian PS. A hybridized ELM for automatic micro calcification detection in mammogram images based on multi-scale features. J Med Syst 2019;43(07):183
84 Chen X, Zargari A, Hollingsworth AB, Liu H, Zheng B, Qiu Y. Applying a new quantitative image analysis scheme based on global mammographic features to assist diagnosis of breast cancer. Comput Methods Programs Biomed 2019;179:104995
85 Arevalo J, Gonzalez FA, Ramos-Pollan R, Oliveira JL, Guevara Lopez MA. Convolutional neural networks for mammography mass lesion classification. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:797–800
86 Taghanaki SA, Kawahara J, Miles B, Hamarneh G. Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification. Comput Methods Programs Biomed 2017;145:85–93
87 Sert E, Ertekin S, Halici U. Ensemble of convolutional neural networks for classification of breast microcalcification from

mammograms. Annu Int Conf IEEE Eng Med Biol Soc 2017; 2017:689–692
88 Tan M, Qian W, Pu J, Liu H, Zheng B. A new approach to develop computer-aided detection schemes of digital mammograms. Phys Med Biol 2015;60(11):4413–4427
89 Saraswathi D, Srinivasan E. A CAD system to analyse mammogram images using fully complex-valued relaxation neural network ensembled classifier. J Med Eng Technol 2014;38(07):359–366
90 Suh YJ, Jung J, Cho BJ. Automated breast cancer detection in digital mammograms of various densities via deep learning. J Pers Med 2020;10(04):E211
91 Jiao Z, Gao X, Wang Y, Li J. A deep feature based framework for breast masses classification. Neurocomputing 2016;197:221–231
92 Kooi T, van Ginneken B, Karssemeijer N, den Heeten A. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. Med Phys 2017;44(03):1017–1027
93 Duraisamy S, Emperumal S. Computer-aided mammogram diagnosis system using deep learning convolutional fully complex-valued relaxation neural network classifier. IET Comput Vis 2017;11(08):656–662
94 Kurek J, Swiderski B, Osowski S, Kruk M, Barhoumi W. Deep learning versus classical neural approach to mammogram recognition. Bull Pol Acad Sci Tech Sci 2018;66:831–840
95 Jiao Z, Gao X, Wang Y, Li J. A parasitic metric learning net for breast mass classification based on mammography. Pattern Recognit 2018;75:292–301
96 Al-antari MA, Al-masni MA, Park SU, et al. An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network. J Med Biol Eng 2018;38(03): 443–456
97 Niu J, Li H, Zhang C, Li D. Multi-scale attention-based convolutional neural network for classification of breast masses in mammograms. Med Phys 2021;48(07):3878–3892
98 Rajathi GM. Optimized radial basis neural network for classification of breast cancer images. Curr Med Imaging 2021;17(01): 97–108
99 Khan HN, Shahid AR, Raza B, Dar AH, Alquhayz H. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. IEEE Access 2019; 7:165724–165733
100 Medjeded M, Saïd M, Chenine A, Chikh M. A new triplet convolutional neural network for classification of lesions on mammograms. Rev Intell Artif 2019;33:213–217
101 Sun L, Wang J, Hu Z, Xu Y, Cui Z. Multi-view convolutional neural networks for mammographic image classification. IEEE Access 2019;7:126273–126282
102 Yu S, Liu L, Wang Z, Dai G, Xie Y. Transferring deep neural networks for the differentiation of mammographic breast lesions. Sci China Technol Sci 2019;62(03):441–447
103 Tavakoli N, Karimi M, Norouzi A, Karimi N, Samavi S. Soroushmehr SMR. Detection of abnormalities in mammograms using deep features. J Ambient Intell Humaniz Comput 2019;14(05): 5355–5367
104 Jo YY, Choi YS, Park HW, et al. Impact of image compression on deep learning-based mammogram classification. Sci Rep 2021; 11(01):7924
105 Cui Y, Li Y, Xing D, Bai T, Dong J, Zhu J. Improving the prediction of benign or malignant breast masses using a combination of image biomarkers and clinical parameters. Front Oncol 2021; 11:629321
106 Yang J, Li H, Shi N, Zhang Q, Liu Y. Microscopic tumour classification by digital mammography. J Healthc Eng 2021; 2021:6635947
107 Chouhan N, Khan A, Shah JZ, Hussnain M, Khan MW. Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography. Comput Biol Med 2021;132:104318
108 Ragab DA, Attallah O, Sharkas M, Ren J, Marshall S. A framework for breast cancer classification using Multi-DCNNs. Comput Biol Med 2021;131:104245
109 Tsochatzidis L, Koutla P, Costaridou L, Pratikakis I. Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses. Comput Methods Programs Biomed 2021;200:105913
110 Zhang X, Liang C, Zeng D, et al. Pattern classification for breast lesion on FFDM by integration of radiomics and deep features. Comput Med Imaging Graph 2021;90:101922
111 Rehman KU, Li J, Pei Y, Yasin A, Ali S, Mahmood T. Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network. Sensors (Basel) 2021;21(14):4854
112 Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020;6 (10):1581–1588
113 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577 (7788):89–94
114 Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 2019;292(02):331–342
115 Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. Radiol Artif Intell 2020;2(06):e190208
116 Watanabe AT, Lim V, Vu HX, et al. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. J Digit Imaging 2019;32 (04):625–637
117 Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med 2021;27 (02):244–249
118 Tsochatzidis L, Costaridou L, Pratikakis I. Deep learning for breast cancer diagnosis from mammograms-a comparative study. J Imaging 2019;5(03):37
119 Chakraborty DP, Breatnach ES, Yester MV, Soto B, Barnes GT, Fraser RG. Digital and conventional chest imaging: a modified ROC study of observer performance using simulated nodules. Radiology 1986;158(01):35–39
120 Uzun Ozsahin D, Ikechukwu Emegano D, Uzun B, Ozsahin I. The systematic review of artificial intelligence applications in breast cancer diagnosis. Diagnostics (Basel) 2022;13(01):45

